White Paper on the Nature and Scope of Issues on Adoption of Model Use Acceptability Guidance

Prepared by

The Science Policy Council
Model Acceptance Criteria and Peer Review
White Paper Working Group

May 4, 1999

TABLE OF CONTENTS

EXECUTIVE SUMMARY

1. What is the purpose of this white paper?

The initial Coordinating Committee to the Science Policy Council (SPC) proposed an examination of options for implementing Agency Task Force on Regulatory Environmental Modeling (ATFERM) recommendations on model acceptance criteria (MAC) and peer review. The Science Policy Council Steering Committee (SPC-SC) accepted this proposal as a means of providing background for decisions on SPC direct involvement in the Models 2000 efforts and the proposed Committee on Regulatory Environmental Modeling (CREM). An Agency-wide task group was formed to consider the options for implementing the ATFERM proposal, and to publish the findings in this white paper.

2. How are the options for interpreting and executing ATFERM recommendations affected by current developments in modeling?

In reviewing current Agency practices in model evaluation and peer review, including evaluating several case histories, the task group observed that **models are more diverse and complex** than in 1993 when the ATFERM final report was written. Therefore, the task group determined that the adequacy of the ATFERM criteria should be reexamined. In the ATFERM report, environmental models were defined in terms of fate and transport, estimation of contaminant concentrations in soil, groundwater, surface water, and ambient air in exposure assessment. Current models used by the Agency range from site-specific to regional in scale; from single pathway and contaminant to multi-pathway and multi-contaminant in operational scope; and from estimating simple exposure results to providing input to complex risk assessments or comparison of management options in function (e.g., "model systems" with component modules and even algorithms uniquely assembled only at the time of application). In 1993, model evaluation and selection were largely Agency functions, whereas inter-agency efforts to develop and use shared models are now more common.

3. What implementation option is recommended to the SPC?

The SPC should engage in direct interaction with the CREM to provide updated general guidelines on MAC to maintain consistency across the Agency (see Section 3.1 for a discussion of other potential options). **Guidelines were recommended as a substitute for criteria** since guidelines would not seem overly bureaucratic to Agency staff or expose the Agency to unnecessary legal challenges regarding model use, but would promote consistency in model evaluation and selection across the Agency. Choices of models to use for environmental decision-making would be left to the program managers, recognizing that model acceptability is related to the specific use of the model and the acceptability of the risk in decision-making due to uncertainty and variability in model inputs. Program managers would be responsible for providing accessible documentation evaluating any model they use. It is anticipated that eventually the CREM would set up a process for periodic review of selected models to provide feedback to Agency senior management on overall consistency of response to the general guidelines.

4. How should the general guidance be developed?

Guidelines should be developed for the various types of peer review that will be used by the Agency in its three-part assessment of models: (1) definition of the objectives, (2) analysis of model uncertainty, and (3) overall assessment. The MAC need to reflect the "state of the art" and be incorporated into an Agency-wide model evaluation strategy that can accommodate different model types and their uses. Heretofore, evaluation criteria did not set explicit specifications that a model must achieve to be suitable for an application. In an integrated assessment of model uncertainty, it is important that explicit specifications be set by program managers for each element of uncertainty (both qualitative and quantitative specifications). The development of element specifications may be influenced by the need to allow for variations in the overall approach, complexity, and purpose of models used by EPA. Using standard model evaluation elements to determine how a model should be assessed could provide a comprehensive integration of the specific model evaluation components into a framework for judging what constitutes a valid model.

5. What follow-up activities should be pursued?

Additional support work is needed in the following areas: 1) analysis of uncertainty, 2) model inventory, 3) multi-media and multi-contaminant model evaluation, 4) comparability of evaluation standards between models.

Suggested follow-up actions include: 1) determining the form, resources needed, and housing of CREM, 2) directing CREM's work toward issuing guidance on "how" to evaluate and characterize models to support the strategy for model evaluation, as opposed to only listing "what" to do, 3) developing and utilizing a model clearinghouse with information on model evaluation results, availability and applications experience, 4) integrating peer review guidance and supporting aspects of QA/QC, 5) preparing case studies to serve as examples of how models used in regulatory decision-making can be evaluated and the value added by such evaluations, and 6) producing a glossary for "state of the art" general guidance to clarify model terminology.

6. What are the roles of QA and peer review, and will a "clearinghouse" be developed?

QA and peer review requirements are imposed to avoid modeling errors that could result in costly mistakes. According to the SPC Peer Review Handbook (EPA 100-B-98-001), models generally should be peer reviewed. Peer review provides an expert and independent third party evaluation that cannot be provided by stakeholder or public comment.

SPC interpretation of ATFERM recommendations would help to clarify what model evaluation records are needed (e.g., code verification, testing results, model selection and the application process). The model evaluation strategy proposed in this document could provide a process tailored to the nature of the predictive task and the magnitude of the risk of making a wrong decision. The proposed strategy could also clarify the complementary roles of QA and peer review tasks in model evaluation and the basis for guidance on QA Project Plans for model development and application.

It is also recommended that the creation of an Agency-wide clearinghouse for models be investigated, since it would provide a means to access model evaluation information while leveraging resources of single organizations and avoiding duplication of effort. Responding to a need perceived by the participants of the 1997 Models 2000 Conference, an action team was formed to consider the development of a Modeling Clearinghouse. To meet the Paperwork Reduction Act of 1980 and OMB Circular A-130 requirements, the Office of Information Resources Management (OIRM) is proposing to develop an Application Systems Inventory (ASI) as a repository of information about Agency software. Another effort, by ORD's National Center for Environmental Assessment, is defining metadata for models that can be stored in its relational database (the Environmental Information Management System) with input through the internet and retrieval through a search engine using specific queries. In addition, a strategy for communication needs to be developed for the public and others, like model users, to provide feedback to the EPA, possibly through the Internet at sites providing information on models and their evaluation for EPA use.

7. What are the benefits and related costs involved in model evaluation?

Evaluation of models during development and peer review would incur costs, but would result in better products. In broader terms, evaluation would also promote systematic management of model development and use within EPA by providing a basis for consistent model evaluation and peer review. The proposed model evaluation strategy would encourage sensitivity and uncertainty analyses of environmental models and their predictions as well as clarify peer review requirements. Access to evaluation results would offer opportunities for improved model selection, elimination of redundant model development and evaluation, and enhanced Agency credibility with external stakeholders. Evaluation and peer review of model application would offer feedback to model developers, hopefully resulting in improved model performance.

1.
INTRODUCTION

1.1 Context

For many years, the Science Advisory Board (SAB) has been actively advising the Agency on the use of computer models in environmental protection. In 1989, after reviewing several models, SAB offered general advice in its first commentary or resolution (EPA-SAB-EEC-89-012), recommending that "EPA establish a general model validation protocol and provide sufficient resources to test and confirm models with appropriate field and laboratory data" and that "an Agency-wide task group to assess and guide model use by EPA should be formed."

In response, the Assistant Administrator for Research and Development (ORD) and the Assistant Administrator for Solid Waste and Emergency Response (OSWER) jointly requested the Deputy Administrator, as the chair of the former Risk Assessment Council (RAC), to establish a task force to examine the issues. The voluntary Agency Task Force on Environmental Regulatory Modeling (ATFERM) was created in March 1992 and completed a report in October 1993. In its report (EPA 500-R-94-001), the ATFERM noted that the Agency has no formal mechanism to evaluate model acceptability, which causes redundant inconsistent evaluations as well as uncertainty about acceptability of models being applied and the results that the Agency obtains with the models. The ATFERM report recommended establishment of acceptability criteria because "a comprehensive set of criteria for model selection could reduce inconsistency in model selection" and "ease the burden on the Regions and States applying the models in their programs". For Section II of their report, they drafted a set of "acceptability criteria." In Section III, they provided the "Agency Guidance for Conducting External Peer Review of Environmental Regulatory Modeling," which was later issued in July 1994 by EPA's Deputy Administrator on behalf of the Science Policy Council (SPC) as EPA 100-B-94-001. ATFERM also proposed a charter for a Committee on Regulatory Environmental Modeling (CREM) to be created by the Deputy Administrator or the new SPC to carry on work begun by the ATFERM and to provide technical support for model users. This proposal was based on SAB recommendations (SAB-EC-88-040, SAB-EC-88-040A, and SAB-EEC-89-012).

In its peer review of the "Agency Guidance for Conducting External Peer Review of Environmental Regulatory Modeling," SAB heartily endorsed the Agency's general approach to conducting peer review of environmental regulatory modeling (EPA-SAB-EEC-LTR-93-008). The Environmental Engineering Committee (EEC) noted the "most important element to the review process is the verification of the model against available data in the range of conditions of interest" with a discussion of compensating errors and suggested "some guidance needs to be provided as to what constitutes adequate model performance." The report also included specific recommendations on organizational and peer review processes. Later SAB asked to work with the Agency on clarifying its own role, along with other peer review mechanisms, to cover substantially new models, significant adaptations of existing models, controversial applications of existing models, and applications with significant impacts on regulatory decisions (EPA-SAB-EEC-COM-95-005). When implementation of these plans had faltered after several years, SAB urged the Agency to move forward in consolidating its gains in modeling. Their recommendation was echoed in a 1997 external review of "Plans to Address

Limitations of EPA's Motor Vehicle Emissions Model" (GAO/RCED-97-210 p. 10).

In part, as a result of SAB's urging, the Agency conducted the Models 2000 Conference in Athens, GA, in December 1997. Dr. Ishwar Murarka represented the SAB and made a presentation noting the increasing complexity of models. He also stressed the importance of verification and validation issues, sensitivity and uncertainty analyses, intra- and inter-Agency coordination, and the need for a peer review mechanism. Dr. Murarka's bottomline was that new approaches are needed to insure that models are developed, used, and implemented appropriately. The Models 2000 Steering/Implementation Team (SIT) is engaged in an on-going SAB consultation with the Environmental Models Subcommittee on the Agency's modeling efforts that began in May 1998.

Recent discussions have revealed that

1. *the Agency would benefit from an integrated strategy/mechanism for dealing with computer models development and use within the Agency or across agencies.*
2. *the Agency is developing multi-media, multi-pathway models in different program offices for different purposes and SAB has initiated an Advisory on a module of one of the models (i.e., TRIM.FaTE).*
3. *a SAB consultation was requested on the follow-up activities of the Models 2000 workshop, on establishment of the CREM, and on the Agency's goals and objectives in establishing the model acceptability criteria and peer review guidelines.*

1.2 Role of the Science Policy Council

The Science Policy Council (SPC), including its associated Steering Committee (SPC-SC), was established by the Administrator as a mechanism[1] for addressing EPA's many significant policy issues that go beyond regional and program boundaries. They noted that the development and application of environmental regulatory models (ERMs) must be viewed within the larger framework of the risk assessment-risk management (environmental decision-making) paradigm currently utilized by the Agency. Ultimately models can be seen as tools with which risk assessments are performed to support risk management decisions. Therefore, it is critical that the purposes, limitations, and uncertainties inherent in an environmental model be understood by the risk assessor applying the model to a risk concern and the risk manager who depends upon the outputs from a model in decision-making. They also need assurance that the model is being utilized consistently across the Agency. Further, it is vital that the process by which a model is developed, the criteria for evaluating its credibility (mathematical validity, approximation to field results, application to different scenarios, etc.) be accessible to the outside world for objective analysis (e.g., external peer review), and to assessment

[1] As such, its goal is the integration of policies that guide Agency decision-makers in their use of scientific and technical information. The SPC works to implement and ensure the success of initiatives recommended by external advisory bodies such as the National Research Council (NRC) and SAB, as well as others such as the Congress, industry and environmental groups, and Agency staff.

by the public[2] at large. Also as modeling becomes more sophisticated and bridges multi-media, multi-pathways, multi-endpoints, it demands a higher degree of technical expertise and training from the EPA.

The recently issued SPC Peer Review Handbook (USEPA, 1998) incorporates the earlier ATFERM guidance on the peer review of ERMs (on the EPA website http://www.epa.gov/ORD/spc). The ATFERM guidance states that "...environmental models...that may form part of the scientific basis for regulatory decision-making at EPA are subject to the peer review policy...and...this guidance is provided as an aid in evaluating the need and, where appropriate, conducting **external** peer review related to the development and/or application of environmental regulatory modeling." The guidance further defines what is encompassed by peer review of model development[3] and applications.[4]

The guidance describes the steps in the external peer review process, the mechanisms, general criteria, and documentation for conducting external peer review and the specific **elements** of peer review. The elements include: 1) model purpose, 2) major defining and limiting considerations, 3) theoretical basis for the model, 4) parameter estimation, 5) data quality/quantity, 6) key assumptions, 7) model performance measures, 8) model documentation including users' guide, and 9) a retrospective analysis. The guidance does not specifically address externally funded model peer review, but Agency policy[5] being developed for implementation of the Cancer Guidelines should provide a useful precedent.

The Risk Assessment Forum (RAF) was established to promote scientific consensus on risk assessment issues and to ensure that this consensus is incorporated into appropriate risk assessment guidance. RAF recently convened a workshop on Monte Carlo analysis (EPA/630/R-96/010), and acting upon the recommendations from the workshop, developed a set of guiding principles to guide agency risk assessors in the use of probabilistic analysis tools. The tools were also provided to support adequate characterization of variability and uncertainty in risk assessment results (e.g., sensitivity analyses). Policy on acceptance of risk assessments was also developed. It requires that the methods used be documented sufficiently (including all models used and all data upon which the assessment is based and all assumptions impacting the results) to allow the results to be independently reproduced.

---

[2] Stakeholder involvement (i.e., involvement by those interested or affected entities) in the development of ERMs within the environmental decision-making framework is both desirable and necessary. It is **desirable** because often the regulated industries or other affected groups have special insight or expertise into parameters, e.g., the industrial process, exposure issues, place or media-based concern, which must be integrated into the EPA model. Their participation provides a value-added dimension to the development process and enhances the chances of model acceptance and/or public credibility. It is **necessary** for the obvious reason of foregoing possible lawsuits and because the public is insisting upon a greater involvement in the decision-making process (Presidential Commission, 1997; NRC, 1996).

[3] Models developed to support regulatory decision-making or research models expanded to develop scientific information for Agency decision-making would be subject to the peer review guidance.

[4] Normally, the first application of a model should undergo peer review. For subsequent applications, a program manager should consider the scientific/technical complexity and/or novelty of the particular circumstances as compared to prior applications. **Peer review of all similar applications should be avoided because this would likely waste precious time and monetary resources while failing to provide the decision-maker with any new relevant scientific information.** Nevertheless, a program manager may consider conducting peer review of applications upon which costly decisions are based or applications which are likely to end up in litigation.

[5] The specific details are not yet available, but external stakeholder groups funding peer reviews of ERMs for Agency use will be expected to generally adhere to the same procedures that the EPA is following.

1.3 Purpose of the White Paper

In follow-up to the Models 2000 Conference, the initial Coordinating Committee to the SPC proposed that an examination of options for implementing the ATFERM recommendations on model acceptance criteria and peer review, and to publish the findings in a white paper. The SPC-SC accepted the proposal as a means of providing background for decisions on SPC direct involvement in the Models 2000 effort and the proposed CREM. An Agency-wide task group representing the National Program offices, Regions, ORD and EPA's Quality Assurance Division was assembled (see Appendix A) to consider whether the ATFERM proposal should be carried out and, if so, with what criteria.

For further clarification, the following questions raised by the ad hoc Coordination Committee are answered (Section 6):

1. How do the issues of Peer Review (external/internal) and QA/QC evaluation relate to acceptability determination?
2. What is a consensus definition of model use acceptability criteria?
3. Does acceptability correspond to a particular model, or specific applications of a model?
4. Does acceptability cover only models developed by EPA or can it cover externally developed models?
5. Does acceptability mean the agency will develop a "clearinghouse" of models that meet EPA's definition of acceptable?
6. Would each program/region develop their own system for evaluating acceptability?
7. Should EPA apply a generic set of criteria across the board to all categories of environmental regulatory models (ERMs) or should acceptability criteria differ depending on the complexity and use (e.g., screening vs. detailed assessment) of a model?

2.
BACKGROUND

2.1 Current Practices in Model Evaluation

The task group first wanted to establish a background in order to address whether or not model acceptance criteria (MAC) should be adopted, and if so, how to define them. They wanted to know if the context for the MAC had changed. Available information focusing on the last five years since the ATFERM report was written, along with case histories in model evaluations (Appendix B), are summarized in the following sections:

2.1.1 Office of Air and Radiation (OAR)

1) The Office of Air Quality Planning and Standards (OAQPS)

OAQPS supports implementing the Clean Air Act (CAA) air quality modeling requirements, which includes several mechanisms to assist the Regional Offices and state and local air pollution control agencies in approving and/or developing models and modeling techniques for air quality dispersion applications. It has weathered the test of time and continues to meet the Regions' needs, therefore recommendations to change its current emphasis or mode of operation are not needed. The implementation process includes the following:

a) Appendix W to Part 51: Guideline on Air Quality Models (Guideline) of 40 Code of Federal Regulations

The Guideline promotes consistency in the use of air quality dispersion modeling within the air management programs. It also recommends air quality modeling techniques that should be applied to State Implementation Plan (SIP) revisions and new source reviews, including prevention of significant deterioration (PSD) permits. A compendium of models and modeling techniques acceptable to EPA is provided. The recommended models are subjected to peer scientific review and/or a public comment and review process. The Guideline specifically addresses the use of alternative models or techniques, if an EPA preferred model or procedure is not appropriate or available.

Several revisions to the guideline have occurred over the years. New modeling paradigms or models proposed for the Guideline are required to be technically and scientifically sound; undergo beta testing, model performance evaluation against applicable EPA preferred

models, and use of field studies or fluid modeling evaluation; documented in a user's guide within the public domain; and undergo some level of peer review. The conferences and modeling revisions are announced and proposed through the Federal Register. Modeling revisions then become a part of the regulatory process after publication of a final notice that addresses public comments and EPA's responses. Many new modeling techniques have successfully made the transition to some level of Guideline acceptance.

b) Model Clearinghouse

Section 301 of the CAA requires a mechanism for identifying and standardizing inconsistent or varying criteria, procedures, and policies being employed in implementing and enforcing the CAA. The Regions are responsible for ensuring that fairness and uniformity are practiced. The Model Clearinghouse was created many years ago to support these requirements. It is a Regional resource for discussing and resolving modeling issues and for obtaining modeling guidance and modeling tools. A primary purpose of the Model Clearinghouse is to review the Regional Offices' positions on non-guideline models or alternative techniques. The Clearinghouse reviews each referral for national consistency before final approval by the Regional Administrator. This requires an historical cognizance of the Clearinghouse on usage of non-guideline techniques by the Regional Offices and the circumstances involved in each application.

In FY-1981, the Clearinghouse began to maintain paper files of referrals from the Regional Offices. These files document the usage of non-guideline models and alternative techniques. The information in the files is summarized and communicated to the Regional Offices periodically to increase awareness of any precedents when reviewing state or industry proposals to apply non-guideline models or alternative techniques.

After a few years, the Model Clearinghouse Information Storage and Retrieval System (MCHISRS) was designed. This is a database system to manage information about Regional inquiries involving the interpretation of modeling guidance for specific regulatory applications. The computer database was recently placed on the SCRAM BBS for wider dissemination and access. The MCHISRS includes key information involving States, pollutants, models, terrain type, and so on, plus a narrative summary. The summary includes a statement of the modeling or modeling related issue involved and the Clearinghouse position on the issue. Any users can now examine the

historical records to determine impact on their particular application.

The Clearinghouse is always accessible to the Regions. A mechanism for timely review of unique modeling applications or nonguideline modeling techniques with a view toward nationwide precedent and consistency is professionally achieved with the Clearinghouse.

c) Support Center for Regulatory Air Modeling Internet Web site, (SCRAM Internet Web site)

An important by-product of the Model Clearinghouse is the development and maintenance of the SCRAM Internet website. This bulletin board has been in existence and accessible to the public for at least a decade. The SCRAM Internet website represents an electronic clearinghouse for the Agency and provides a critical linkage to the Regions and the public. It maintains a historical record of guidance on generic issues and other policy memoranda and is updated as needed. All of the Agency's models and user's guides that are recommended for regulatory use are maintained through OAQPS and available at all times through the SCRAM Internet website.

In addition to the Agency preferred models, a variety of alternative models and support computer programs are available through SCRAM Internet website. This Internet website also provides complete and timely documentation of not only the revisions to these models but documentation on why they were needed and their effects on model performance. Although the basic format of the Internet website has not changed significantly, changes are made to better meet the needs of the customers and the ever broadening scope of air dispersion modeling.

The recent move of the bulletin board to the Internet is just one example of how OAQPS works to improve accessibility of this system. The SCRAM Internet website is one of the most user-friendly bulletin boards on the INTERNET. It appears that the majority of the Regions' needs that are related to the successful implementation of the CAA air quality dispersion modeling regulations and requirements are met by the Clearinghouse and the SCRAM Internet website.

d) Conferences on Air Quality Modeling

Section 320 of the CAA requires EPA to conduct a conference

on air quality modeling at least every three years. The Seventh Modeling Conference on Air Quality Modeling should be held this year. The conference provides a forum for public review and comment on proposed revisions to the Guideline.

e) Periodic Modeling Workshops

Finally, an annual workshop is held with the EPA Regional Meteorologists and state and local agencies to ensure consistency and to promote the use of more accurate air quality models and databases for PSD and SIP-related applications.

2) Office of Radiation and Indoor Air (ORIA)

Intra- and Inter-Agency cooperative efforts developed technical guidance on model selection and evaluation through a joint Interagency Environmental Pathway Modeling Working Group. The group was established by the EPA Offices of Radiation and Indoor Air (ORIA) and Solid Waste and Emergency Response (OSWER), the Department of Energy (DOE) Office of Environmental Restoration and Waste Management, and the Nuclear Regulatory Commission (NRC) Office of Nuclear Material Safety and Safeguards. Their purpose was to promote more appropriate and consistent use of mathematical environmental models in the remediation and restoration of sites contaminated by radioactive substances.

First, the EPA, DOE, and NRC working group sponsored a mail survey in 1990 and 1991 to identify radiologic and non-radiologic environmental transfer or pathway computer models that have been used or are being used to support cleanup of hazardous and radioactive waste sites. The intent of the survey was to gather basic administrative and technical information on the extent and type of modeling efforts being conducted by EPA, DOE, and NRC at hazardous and radioactive waste sites, and to identify a point of contact for further follow-up. A report, Computer Models Used to Support Cleanup Decision-Making at Hazardous and Radioactive Waste Sites, was published (EPA 402-R-93-005, March 1993) to provide a description of the survey and model classification scheme, survey results, conclusions, and an appendix containing descriptions and references for the models reported in the survey.

Later, reports resulting from the working group's efforts were published (described below) to be used by technical staff responsible for identifying and implementing flow and transport models to support cleanup decisions at hazardous and radioactive waste sites. One report, Environmental Pathway

Models: Ground-Water Modeling in Support of Remedial Decision-Making at Sites Contaminated with Radioactive Material, (EPA 402-R-93-009, March 1993) identified the role of, and need for, modeling in support of remedial decision making at sites contaminated with radioactive materials. It addresses all exposure pathways, but emphasizes ground-water modeling at EPA National Priority List and NRC Site Decommissioning Management Program sites. Its primary objective was to describe when modeling is needed and the various processes that need to be modeled. In addition, the report describes when simple versus more complex models may be needed to support remedial decision making.

A Technical Guide to Ground-Water Model Selection at Sites Contaminated with Radioactive Substances (EPA 402-R-94-012, September 1994) was prepared to describe methods for selecting ground-water flow and contaminate transport models. The selection process is described in terms of matching the various site characteristics and processes requiring modeling and the availability, reliability, validity, and costs of the computer codes that meet the modeling needs.

Another report, Documenting Ground-Water Modeling at Sites Contaminated with Radioactive Substances, (EPA 540-R-96-003, January 1996) provided a guide to determining whether proper modeling protocols were followed, and, therefore, common modeling pitfalls avoided. The problems were noted in a review of 20 site-specific modeling studies at hazardous-waste remediation sites (Lee et al., 1995). The review cited problems in 1) misunderstanding of the selected model, 2) improper application of boundary conditions and/or initial conditions, 3) misconceptualization, 4) improper or unjustifiable estimation of input data, 5) lack of or improper calibration and verification, 6) omission of or insufficient sensitivity and uncertainty analysis, and 7) misinterpretation of simulation results. Any of these errors could impact remedial and risk decisions. As a guide to modelers, this report demonstrates a thorough approach to documenting model applications in a consistent manner. A proper documentation of modeling results was found to answer the following questions:

- Do the objectives of the simulation correspond to the decision-making needs?
- Are there sufficient data to characterize the site?
- Is the modeler's conceptual approach consistent with the site's physical and chemical processes?
- Can the model satisfy all the components in the conceptual model, and will it provide the results necessary to satisfy the study's objectives?
- Are the model's data, initial conditions, and boundary conditions identified and consistent with geology and hydrology?

- Are the conclusions consistent with the degree of uncertainty or sensitivity ascribed to the model study, and do these conclusions satisfy the modeler's original objectives?

The approach recommended for evaluating models consists of three steps: (1) determining one's objectives and data requirements for the project; (2) properly developing a conceptual model for the site, which describes the physical and chemical system that must be simulated; and (3) selecting and applying the model in a manner consistent with the objectives and the site's known physical characteristics and input variables.

3) The Office of Atmospheric Programs

See RADM case history in Appendix B.

4) The Office of Mobile Sources (OMS)

OMS's model evaluation includes extensive stakeholder review, increased amounts of external peer review, and alpha- and beta- testing of models. Recent efforts have included using the ATFERM guidance and the Agency's peer review policies for conducting more extensive peer review and model evaluation. In addition, efforts are underway to determine the best and most efficient way possible to determine uncertainties in models.

2.1.2 Office of Solid Waste and Emergency Response (OSWER)

In 1989, OSWER undertook a study to examine its modeling environment. OSWER staff found more than 310 models in use in the hazardous waste and Superfund programs. Many of the earlier models were written in Fortran. The newer models, many written to run on microcomputers, used a variety of languages and tools. These models varied in their applications and design. Efforts to verify, validate, and select models were inconsistent with little overall guidance and user support. The report concluded with three recommendations:

Task Area 1: Initiation, Additional Study, and Preparation of a Management Plan.

Task Area 2: Development of Guidance for Modeling.

Task Area 3: Establishment of User Support Network for HW/SF Modeling.

This study prompted OSWER's leadership in the development of the

subsequent Agency Report of the Agency Task Force on Environmental Regulatory Modeling (EPA 500-R-94-001) and the Guidance for Conducting External Peer Review of Environmental Regulatory Models (EPA 100-B-94-001).

The situation today has become even more complex with the advent of microcomputers and fourth generation languages that facilitate rapid development of computer programs. However, most of the challenges that faced EPA when the OSWER modeling study was undertaken still exist today. For example, the threat of legal challenge to the use of models for regulatory applications continues.

Recently, a Validation Strategy has been developed for the IEUBK model (EPA/540/R-94-039). The approach emphasizes verification, validation and calibration as previously established through the ATFERM report for environmental exposure models even though the model is for blood lead levels in children rather than exposure assessment. It uses 4 components:

1. **Scientific foundations of the model structure**. Does the model adequately represent the biological and physical mechanisms of the modeled system? Are these mechanisms understood sufficiently to support modeling?

2. **Adequacy of parameter estimates**. How extensive and robust are the data used to estimate model parameters? Does the parameter estimation process require additional assumptions and approximations?

3. **Verification**. Are the mathematical relationships posited by the model correctly translated into computer code? Are model inputs free from numerical errors?

4. **Empirical comparisons**. What are the opportunities for comparison between model predictions and data, particularly under conditions under which the model will be applied in assessments? Are model predictions in reasonable agreement with relevant experimental and observational data?

OSWER believes that at least some of these principles would also work for model applications.

2.1.3. Office of Water (OW)

The Office of Science and Technology (OST) in OW is not currently evaluating any models along the lines of the ATFERM acceptance criteria. However, there are two models that will probably be put through the peer review process in the future. Aquatox, is being internally evaluated and later the model will be peer reviewed using

criteria like those in the ATFERM report. After it has completed development, CORMIX will also be peer reviewed.

Another effort is the review of the Basins version 2, an integrated package of a geographic information system, spatial and environmental data, customized analytical tools, watershed and receiving water quality models, and model post processing to be used in analysis of watersheds and preparation of reports and records for Total Maximum Daily Loads. The review objectives do not address evaluation of the models themselves. Long established models like HSPF and QUAL2E have been tested and peer reviewed in the past. However, past evaluation information may not be accessible (e.g., 1980 tests of HSPF that had 10 year record retention schedules).

2.1.4 Office of Prevention, Pesticides and Toxic Substances (OPPTS)

OPPTS does not have a standard approach to model evaluation. Models have been largely developed by consultants with variable evaluation practices. Also a score of different models are used in OPPTS; they range from the trivial to the very complex and in each case the model evaluation depends on its complexity. For example, recently a large consulting firm developed a model to be used at OPPTS for the Acute Dietary Exposure Analyses and Risk Assessment. This model produces a realistic calculation of dietary exposure and includes a Monte Carlo analysis to estimate dietary exposure from potential residues of the total chemical residues in food. It also uses a huge data base that conveys the food consumption habits in the USA. The primary evaluation of the model was done following the Quality Assurance and Quality Control procedures of the vendor. A second in-house evaluation of the model has been conducted through peer review. Statisticians, managers, scientists, computer programmers, and outside consultants evaluated the algorithms of the model to reach a consensus that the model is correct and closely represents reality. However, no formal structured form of model validation (i.e., a mathematical validation) has been used on this particular model. A field verification of this model is not possible because of lack of data. The model validation process rests heavily on a balance reached through a consensus among the parties involved and a constant flow of information between the vendors, the reviewers, and the users. Ultimately, the Scientific Advisory Panel, an external peer review group mandated by the Federal Insecticide, Fungicide and Rodenticide Act (FIFRA) is responsible for reviewing major new models used by OPPTS.

2.1.5. Office of Research and Development (ORD)

ORD has been involved in model evaluation as well as development in its support for the National Program Offices and Regions. Case histories in model evaluation (Appendix B) demonstrate a wide range of approaches from traditional calibration and validation (SHADE-HSPF and RADM) to benchmarking with other Federal Agencies (MMSOILS). Model systems are also being developed similar to

that planned by OAQPS (TRIM.FaTE). Peer review mechanisms used in the case histories include internal review, review by EPA and non-EPA (e.g., DOE) advisory committees, and journal peer review of articles.

In 1994, a protocol for model validation (in the Draft white paper "Model Validation for Predictive Exposure Assessments" see Appendix C) was prepared at the request of the Risk Assessment Forum. The protocol was developed from a design perspective to provide a consistent basis for evaluation of a model in performing its designated task reliably. A wide variety of evidence was proposed to inform the decision rather than just the conventional test with matching the model output to historical data (history matching). Therefore, the protocol could cover the case where predictions make extrapolations from past observations into substantially altered future situations. The protocol addressed three aspects:

1. The intrinsic properties of the model;
2. The nature of the predictive task; and
3. The magnitude of the risk of making a wrong decision.

It was noted that models would differ in the level of evaluation possible. If the prediction task was routine where quantitative results of model performance could be evaluated, the risk of making a wrong decision would be low. If the prediction task was novel, where little previous experience existed on the model's performance, evidence would be more qualitative (e.g., peer reviewer's opinions on the composition of the model), the risk would be higher.

The types of evidence supporting model evaluation were outlined and included:

A) Structure - the conceptual basis (easier for established theories but hard to quantify uncertainty) and the way in which the constituent model mechanisms (hypotheses) are expressed mathematically and connected to one another;

B) Complexity in number of its state variables and parameters (e.g., ecological models of environmental systems would have more hypotheses and it would be more difficult to detect a failure in any one hypothesis or to predict its impact on the prediction);

C) Values of the parameters and the bands of uncertainty around them (related to the data quality) and extent of the observations;

D) Sensitivity of the outputs of the model to changes in assigned values of each parameter; and

E) History matching with field data which can include quantitative evaluation of model performance if criteria for success are provided.

In modeling supporting regulatory uses, the evaluation process results in a choice to use a model as a tool for prediction. This emphasizes the perspective of the quality of the decision and tests the null hypothesis that the model adequately represented the process modeled until shown to be otherwise. The sum of the evidence

would be used, however, methods for weighting types of evidence are needed. Unfortunately the term of the most knowledgeable advocate of the Risk Assessment Forum ended before action was taken and the other members did not pursue the protocol.

2.2 Summary

The various National Program Offices and the Office of Research and Development vary in their approach and practices in model evaluation as well as in the types of models they use. In our review of program information and model evaluation case histories (Appendix B), we noted that models are becoming more diverse ranging from site-specific to regional in scale; from single contaminant and pathway to multi-pathway and multi-contaminant in operational scope; and from estimating simple exposure results to providing complex risk assessments or comparisons of management options in function. They are also more complex as "model systems" are being developed with component modules and even algorithms that will be uniquely assembled only at the time of application. Inter-agency efforts are also more involved in evaluation and selection of models in shared projects. This situation varies from 1993 when the ATFERM final report was written which defined environmental models in terms of fate and transport, estimation of contaminant concentrations in soil, groundwater, surface water, and ambient air in exposure assessment (page III-1).

3.
OPTIONS FOR SPC INTERACTION WITH CREM

3.1 Options Considered by the Task Group

1. ***Do nothing.*** *This option implies that the current peer review policy, i.e., the SPC (Section III, ATFERM report) referenced in the SPC Peer Review Handbook (USEPA, 1998), would remain the basis for decisions on model evaluation. This guidance recommends external peer review, which remains to be defined as to its precise nature, but would have to have some objective standards/criteria; furthermore, the ATFERM MAC would have to be sorted out as to their appropriateness/present utility for external peer review. This leads us back to the need for generally acceptable MAC for the Agency and to the repeated recommendations of the SAB for a mechanism such as CREM to address Agency models.*

2. ***Establish the CREM and have them responsible for reviewing "environmental" models for acceptability in regulatory use (e.g., policy, regulatory assessment, environmental decision-making) and list acceptable models as proposed in the ATFERM report using criteria as listed in the ATFERM report or a revision of the criteria.*** *The CREM would implement the model evaluation process and upgrade the model information system. This could be accomplished through either the models website or a clearinghouse library that provides information on how models satisfy the model acceptability criteria and access to the models. Model use acceptability criteria as listed in the ATFERM report or a revision of the criteria addressing changes in model evaluation practices and the increased complexity and range of models in use across the Agency, would be used as the standards for information reporting. Generic criteria with specific programmatic refinements (e.g., quantitative elements) could be envisioned.*

3. ***Leave decisions on regulatory use to program managers (who provide specifications like quantitative criteria) and their technical staff but require accessible information responding to the model acceptance criteria.*** *Again, this could be accomplished through either the models website or a clearinghouse library that provides information and access to the models. Model acceptability criteria as listed in the ATFERM report, or a revision of the criteria, addressing changes in model evaluation practices and the increased complexity and range of models in use across the Agency, would be used as the standards for information reporting. Revision of the ATFERM model acceptance criteria would be addressed through a phased evaluation (e.g., development evaluation with qualitative criteria then application with quantitative criteria related to regulatory application) and analysis of the most appropriate kinds of peer review to apply to model development and use.*

3.2 Task Group Recommendations

The task group **recommends a combination of options 2 and 3**. Decisions on regulatory use of models would be left to program managers recognizing model acceptability is related to its specific use, however, the Science Policy Council should engage in direct interaction with the CREM to provide updated general guidelines on model acceptance criteria to maintain consistency across the Agency. The program managers should respond to the updated general guidelines on model acceptance by developing specifications related to the model types and use in their programs and assuring model information responding to the criteria is accessible. Model acceptance criteria will help define general acceptability for model developers as well as assist users to select and apply models appropriately. The CREM should provide feedback to Agency senior management on consistency in response to the general guidance after periodic review of selected models.

4.
PROPOSED SCOPE, APPROACH, AND SUPPORTING ANALYSIS FOR THE GENERAL GUIDANCE

4.1 Scope ! Handling Uncertainties in Model Evaluation

Model evaluation must address both qualitative and quantitative uncertainties (Beck et al. 1997). **Qualitative uncertainty** arises in the comparative analysis of the model's structure to the environmental component addressed in the regulatory task. Structure is the way the constituent components or hypotheses are expressed mathematically and connected. Each hypothesis can be judged untenable if model predictions are found to be incompatible with observations of reality. Finding invalid components or connections is more difficult in complex models. Evaluation of the key model components and their redundancy can help us discriminate between a match and a mis-match in qualitative observed behavior and the model structure. However, it is difficult to quantify the impact on results in predictions from structural errors.

Quantitative uncertainty occurs in values assigned to model coefficients (parameterization) and is related to the amount and quality of relevant data (e.g., variation of contaminant concentration in time and space) available. Matching of the model's predictions (performance) to past observations, even when reasonable agreement is found, can mask real uncertainty in the model's approximation of the environment's behavior. Mis-characterization of one parameter can mask, or compensate for mis-characterization of one or more other parameters. So evaluation of the uncertainty in calibration of the parameters should be quantified (e.g., variances and covariances of parameter estimates or bands of uncertainty) to support model selection (e.g., match the regulatory task alternatives to a model only in areas where they are maximally insensitive to the known uncertainties). Thus a strategy can be developed identifying the objective evidence to be considered in model evaluation and how to represent the weight of evidence in the model's success or failure to perform its designated task.

The overview of model evaluation above, however, does not address all problems that occur in the site specific application phase. Agency program managers need to need to be aware that other factors may also affect the viability of model predictions. Model outputs or predictions must be evaluated to identify erroneous and uncertain results from improper input data, improper boundary condition specification, unjustified adjustment of model inputs as well as violations of model assumptions and exercising of the model outside its proper domain and application niche (EPA-SAB-EEC-LTR-93-008).

Although peer review is mentioned as a touchstone in the model evaluation process by both the ATFERM report and Beck et al. (1997), difficulties in applying effective peer review should not be underestimated. First, model evaluation must be conducted for the range of processes from model construction through to the regulatory use of its outputs. For each process different peers may be appropriate. Because of the technical nature of model construction, there is a tendency to focus on model constructors as the principal, if not the only peer reviewers. This can emphasize a journal style

peer review approach, which may be necessary, but is not sufficient according to the Peer Review Handbook.

Second, peer review can rarely make a comprehensive analysis of a model including the technical aspects of its implementation. For example, an essential part of a model code verification includes a line-by-line analysis of computer code, which is not a task that a peer reviewer, in the traditional sense, is able to complete. These particular difficulties, when combined with the general difficulties known to exist with peer review such as apparent subjectivity in qualitative areas (e.g., van Vallen and Pitelka 1974, Peters and Ceci 1982, Cichetti 1991), mean that we can not rest on peer review as though it were a gold standard. Peer review can be useful for some functions but is inadequate for others, such as model code development and the analysis of practical questions about the use of models in a regulatory framework unless the term, "peers", is defined more broadly than ever before. For example, the questions in the peer review charter could address specific model evaluation records such as code verification (if general guidance on the necessary records was provided) and ask reviewers if the evaluation was adequate. The peer review panel would need to be constructed to contain appropriate expertise in this area.

For it to be an effective part of model evaluation, guidelines need to be developed for the types of peer review that should be applied to the different components of uncertainty and precisely how these should be used by the Agency in its three-part assessment. Specifications are needed when peer review is used in the analysis of particular elements. The type of peer review used might range from letter reviews by individual researchers, internal or external to the agency, providing individual opinions to panels of scientists, managers and end users of model outputs making a comprehensive analysis of model development and use. Supplying the technical specifications to reviewers for them to use in assessing particular elements of uncertainty would be an innovation for review of models but it is similar to the technical standards used in manuscript review.

4.2 Approach

4.2.1 Strategy for Model Evaluation

Considering that the Agency's regulatory actions are often challenged, the model acceptance criteria need to reflect the "state of the art" in model evaluation (see Appendix D). The criteria also need to be incorporated into an Agency-wide strategy for model evaluation that can accommodate differences between model types and their uses. As discussed in Section 3.1.5, a protocol was developed for the RAF to provide a consistent basis for evaluation of a model's ability to perform its designated task reliably (see Appendix C). The protocol offers flexibility by proposing a wide variety of evidence to support the decision and covered even situations lacking field and laboratory data. The protocol also includes the nature of the predictive task and the magnitude of the risk of making a wrong decision as context for evaluating the intrinsic properties of the model. Because the evaluation process results in a choice of whether or not to use a model as a tool for prediction, the perspective of the quality of the decision is emphasized. Also, the protocol accepts the null hypothesis that the model

adequately represents the modeled process until shown to be otherwise which is consistent with current thinking. This protocol is a good place to start, however, its terminology needs to be updated.

4.2.2 Strategy for Defining Uncertainty in Model Elements

While there is no clear and universally applied definition of the model evaluation process, there is considerable overlap between suggested strategies and their working definitions. For example, there is repeated use of model evaluation elements despite considerable differences in model types and applications of available evaluation techniques. This informal consensus provides the potential for characterizing model evaluation by analyzing and defining the following elements:

- Uncertainty in the theory on which the model is based.
- Uncertainty in translating theory into mathematical representation.
- Uncertainty in transcription into computer code.
- Uncertainty in assigning parameter values and model calibration.
- Uncertainty in model tests.

This approach could be used to identify those elements that must rely on **peer review**; those that should use **quantitative measures** (e.g., decision maker specifies the acceptable agreement in accuracies and precisions between distributions of model outputs and data from field sampling); and those that should be assessed by an **agency panel** (e.g., users, stakeholders, and program staff to address effectiveness and accessibility of computer code). Whereas modelers and scientists familiar with a model tend to focus on improving particular elements of uncertainty, the EPA requires a comprehensive and consistent basis for evaluating if a model performs its designated task reliably. This approach also offers the potential of a comprehensive integration of the different model evaluation components into a framework for judging what constitutes a valid model in a specific situation rather than the judgement itself (left to the program managers). A wide body of evidence needs to be included in such a process because no evidence is above dispute, even "objective" measures of performance depend on some subjectivity in the chosen level of an acceptable difference between a pair of numbers (quantitative criteria relevant to the program).

4.3 Supporting Analysis

Our synthesis (see Table 4.1) focuses on the ASTM D978-92 process description of model evaluation (but not definition), the RAF protocol discussed as a starting point (similar to discussions in

Beck et al. 1997 and Rykiel 1996), and the ATFERM model acceptance criteria questions. ASTM describes model evaluation in terms of seven processes: conceptualization, program examination, algorithm examination, data evaluation, sensitivity analysis, validation, and code comparison as well as documentation and availability. In the academic literature some authors have focused on different statistics that might be used in model evaluation such as calculations for particular types of variables and for use at different stages in the model evaluation process. However, no definitive protocols have emerged on when or how these statistics should be used. The RAF protocol considers: structure, complexity in number of its state variables and parameters, values of the parameters and the bands of uncertainty around them, sensitivity of the outputs of the model to changes in assigned values of each parameter, and history matching with field data which can include quantitative evaluation if criteria for success are provided to evaluate model performance. The ATFERM report suggests four ways to evaluate and rank reliability of models: (1) appropriateness, (2) accessibility, (3) reliability and (4) usability and lists questions to be asked under each heading but provides no specifications for answering them.

TABLE 4.1 The ASTM E978-92, RAF Protocol, and ATFERM report each identify some of the same issues in model evaluation giving them different names. Nomenclature used in ASTM E978-92, left column, is compared with the other two reports in the next two columns, respectively.

ASTM E 978-92	RAF Protocol	ATFERM
(A^3) Model Conceptualization	Structure & Composition (5.1.1i)[1]	(3a)[2] Theoretical basis peer reviewed?
(A, B) Program Examination	Task Specification (5.1.3) Decision Risk (5.1iii)	(1a) Application niche? Questions answered?
(B,C) Algorithm Examination	Mathematical Expression of hypotheses & Complexity (5.1.1ii)	(3b) - algorithm & code peer review; (4bvii) code verification
(C, D) Data Evaluation	Parameterization, number & uncertainty (5.1.1 iii, iv, v & vi)	(4d) adequate data availability?
(D) Sensitivity Analysis	Parameter Sensitivities (5.1.1vii)	
(E) Validation	Performance - sequence of errors; paired/unpaired tests; calibration; prediction uncertainty (5.1.2 i-v)	(1b) Strengths, weaknesses, & applicability relative to application niche? (3defg) testing against field data? user acceptability? Accuracy, precision & bias? Code performance?

(E) Code Comparison with similar models		(3 c) Verification testing against accepted models?
Model Documentation		(4b) Full documentation?
Model Availability		(2a) Accessibility? Cost?
C		(4a) code structure and internal documentation?
C		(4c) user support?
C		(4egh) pre- & post-processors? required resources?

[1,2] Numbers correspond to those in the Appendix C protocol and 1994 ATFERM Report EPA 500-R-94-001 questions, respectively. [3] The letters correspond to the steps in Part II of the strategy described below.

The generic guidelines for model evaluation (e.g., ATFERM and ASTM guidelines above) are constructed as a series of questions (e.g., "How well does the model code perform in terms of accuracy, bias and precision?"). They do not set explicit specifications that a model must reach before it is suitable for application (e.g., two or more simulations with acceptable quantitative results over a defined range of scenarios). In an integrated assessment of model uncertainty, it is important that explicit specifications should be set by program managers for each element of uncertainty (both qualitative and quantitative specifications). This would allow flexibility to cover variation in the overall approach, complexity, and the purpose of models used by EPA that may influence the development of such specifications.

The task group's recommendation is to be consistent and use the SPC definition of uncertainty in the sense of "lack of knowledge about specific factors (coefficients), parameters (data), or models (structure)" (EPA/630/R-97/001, 1997 page 8) in the following categories and specify how it is to be assessed. It is recommended that the strategy include three parts:

4.3.1 Part I - Defining the Objectives for the Model

This part would describe the required task for the model, its application niche (ATFERM 1.a), and the precise type of use that is required, (e.g., exactly how it is envisaged that the model would be used in management and/or regulatory tasks and decision making). This part is essential to set the stage for providing detailed answers in Part II which in turn should lead to Part III, a comprehensive evaluation of model uncertainty. In some instances complex conceptual models have been built over years of investigative science that have been combined with predictive modeling. In other cases new conceptual models must be developed as very new problems arise and both researchers and program managers may use a simple approach in prediction. Neither complexity nor simplicity can be judged as "correct," both may have their place, and we require standards for evaluating both. Also, different types of models have very different types of output. A particularly important distinction is between models that have one or more of their major outputs continuously (or at least regularly) measured contrasted with models that do not. In the first case, there may be a useful measure against which to evaluate the model, while in the later case, evaluation may have to be indirect and using a range of different measurements.

4.3.2 Part II - Analysis of Model Uncertainty

The following elements are suggested for evaluating model uncertainty.

A. Uncertainty in the theory on which the model is based.

An appropriate theory must underlie the model. Alternative theories must have been considered and rejected on scientific grounds, and a procedure must be specified for the conditions when new findings will either influence

model structure or cause the development of a new model. Assessing this element of model uncertainty seems most likely to involve peer review and it should be specified how this can be done, e.g., individual reviewers, panels, workshops, and the charters specifying the output from peer review must be explicit. Peer review must also take into account the nature of the task, i.e., that the theory used is relevant to the task.

It is quite likely that different programs of the Agency would place different emphasis on this aspect of uncertainty. For some ecological and environmental systems there is little difference between scientists in views about underlying theory (though there may be substantial differences in how to measure them), in others theoretical differences are important. This corresponds to 3a of ATFERM MAC and Protocol 5.1.1 (i) Structure.

B. Uncertainty in translating theory into mathematical representation.

Alternative mathematical representations of particular functions must be discussed and quantitative evidence given to back particular decisions. Important choices are sometimes influenced by the desire to make a model run more quickly - these must be specified. Error propagation associated with the numerical procedures used to approximate mathematical expressions must be defined. The origin of values for model parameters must be defined and how they were obtained must be described. Assessment guidelines should specify which tasks should be undertaken by peer review and which require a more detailed approach than peer review can provide, and how that will be achieved. This corresponds to Protocol 5.1.1 (ii), (iii),and (iv), and the first part of 3b of ATFERM MAC although not with the same emphasis on peer review.

C. Uncertainty in transcription into computer code.

This stage is required to verify that there are no programming errors (e.g., that the scientific algorithms are properly embedded), that code meets QA/QC specifications and adequate documentation is provided. Programmers are frequently required to make choices in implementing the algorithms that can lead to uncertainty. These need to be documented. Also the code must pass performance tests (e.g. stress tests - NIST Special Publication 500-234). Guidelines must specify how, and by whom, this would be done and if not done by Agency personnel, provide for acquiring the test and documentation records. It may also be necessary to specify programming languages. This corresponds to the second part of 3b of ATFERM MAC and 4a through 4h.

D. Uncertainty in model calibration.

There are three aspects:

(a) uncertainty in the data used in calibration. This should cover everything from numbers of variables to measurements and sampling intensity.

(b) uncertainty in the techniques employed in the calibration procedure. This should cover the use of different optimization techniques and requirements for their implementation.

(c) uncertainty in the parameters obtained. Of course these uncertainties are the result of many of the previous uncertainties. But there should be explicit assessment of parameter ranges.

Guidelines for assessment of these uncertainties should specify what are satisfactory sources of data and sets of calibration procedures. There are likely to be substantial differences between areas of modeling in the assessment of calibration uncertainty. But this type of uncertainty is likely to be an important focus of any external critique and so should be addressed specifically. ATFERM MAC 3c through 3g refer to some of these points but does not make the distinction between calibration and testing. **Sensitivity Analysis** should be contained under this type of uncertainty since it shows the range of potential outputs from the model under different types and/or levels of assumptions, e.g., Protocol 5.1.1 (vii).

E. Uncertainty in model tests.

As with uncertainty in model calibration there are four aspects:

(a) quantity of data available to make test(s),

(b) uncertainty in the data used in making the test(s)

(c) the range of statistics (number and types of tests) to use in any assessment, and

(d) uncertainty in how a test will be actually made (e.g., how is a difference between a calibration and a test to be assessed?).

These points influence the power of any tests and therefore the effectiveness of the assessment that can be made. Protocol 5.1.2 gives examples of particular types of tests, (e.g., unpaired and paired tests). However, the types of tests that can and should be used are perhaps the most variable thing between different types of models, at least at present and they are likely to remain items for considerable research for example, the value of using multiple assessment criteria rather than just one, and how a multi-criteria assessment can be represented.

4.3.3 Part III - The Overall Assessment

This part should refer back to the purpose of the model development. Does the model do the required task? The issue is raised in the Protocol of the magnitude of the risk in making a wrong decision, see Protocol 5.1.3. Applications of the model provide "tests" of its strengths and weaknesses. If adequate documentation is recorded, programmatic evaluations can be conducted to identify areas needing corrective action (e.g., Lee et al, 1995 discussed on p.14). Some of the National Program Offices have developed guidelines for model evaluation for their regulatory applications and these provide examples of the types of specifications that should be expected for each element of uncertainty. For example, OAQPS has developed a consistent practice for air quality models to assure the best model is used correctly for each regulatory application (Guideline on Air Quality Models, 40 CFR, Ch. 1, pt. 51, App. W, 7-1-97). In Section 3 criteria are presented for EPA's evaluation of new models including quantitative specifications for equivalency of new models (e.g., demonstrated by producing maximum concentration predictions within 2 percent of estimates from preferred models). Currently "best estimates" provided by modelers are used (Section 10) noting that errors in the magnitude of the highest estimated concentration occurring sometime within an area of 10 to 40 percent are typically found but acceptable because they are within the factor-of-two accuracy.

5.
POSSIBLE FOLLOW-UP ACTIONS

5.1 Additional Support Work Needed

Research is increasing in the development of techniques appropriate for analysis of the different elements of uncertainty. These techniques could address specific types of model development and application. Some research and development work is needed to support these recommendations:

a) Analysis of Uncertainty

The analysis of uncertainty provides a unifying strategy for model evaluation across the different National Program Offices and Regions of the Agency. However, uncertainty is used in many groups of scientists, e.g., fuzzy set theorists, statisticians, resource managers and risk analysts. A review needs to be made of these uses -- if only to document where we may stand relative to others. Methods of quantifying and combining measures of model uncertainties (e.g., to quantify the results in predictions from structural errors) need to be developed along with modes of presentation to assist decision makers in interpreting uncertainty in the context of regulatory use.

b) Model Inventory

A more comprehensive inventory of the types of models actually used in EPA is needed. Such an inventory might be undertaken by program offices and ORD and incorporated into the Application Systems Inventory proposed by the Office of Information Resources Management's Enterprise Information Management Division to comply with A130. From the last inventory in 1995 and later work by ORD the types of models to be covered are presented in Appendix E.

c) Multi-media and Multi-contaminant Model Evaluation

Multi-media and multi-contaminant models composed of modules grouped together in different ways to answer specific questions pose a complex problem for model evaluation. Have the modules been evaluated for all possible conditions under which they may be used? Should they be?

d) Comparability of Evaluation Standards Between Models

The issue of tailoring the specifications for evaluation to the model's phase of development, complexity, and the types of model structure, needs to be analyzed. Frequently an environmental problem can be pressing, yet there is little information (data) about it. Models may be one of the few ways of analyzing and estimating effects. Obtaining an approximate estimate of an environmental process with an initial model may be an important contribution (see EPAMMM case history in Appendix B). However, there may not be a clearly defined assessment procedure for the model, particularly in comparison to more well established models. How can differences in the specifications of model evaluation be justified and how can an effective model development and its evaluation procedure, be charted through such stages?

5.2 Suggested Follow-up Actions

1. ***Determining the form, resources needed and appropriate housing of the CREM.*** The overall recommendation of the SPC Steering Committee was that the proposed CREM provide updated guidance for the Agency. Thus CREM might be viewed as analogous to the EMMC as a cross-Agency effort to coordinate and promote consistency in model evaluation and use. This presumes that the goal of developing a consistent approach for modeling issues is desirable, if not essential, to the EPA's modeling programs. Although beyond the scope of this paper, it is anticipated that the Models 2000 SIT will present to the SPC a proposal for a charter specifying the CREM's general function, projected resource requirements and structural placement within the Agency in conjunction with the white paper recommendations. In the future, a careful economic assessment by a contractor of the needs of each Program office and the Regions would be valuable since only limited information on model assessment is currently available. In addition, it has been suggested that the CREM might be housed under the auspices of the Risk Assessment Forum.

2. ***Directing the CREM's work toward issuance of peer-reviewed guidances on "how" to evaluate and characterize models to support the strategy for model evaluation (Section 5) rather than only listing "what" should be done***. A rough five year time frame for these guidances is estimated. Examples of guidance subjects needed are:

   - appropriate methods for peer review of models to address uncertainty in model theory;
   - mathematical approaches to code verification to address uncertainty in transcription into model code;
   - the appropriate use of sensitivity analyses in modeling to address uncertainty in model calibration;
   - characterizing applicability of a particular model (needed data, driving parameters, responsiveness to data, etc.) to address uncertainty in model tests and the overall assessment with respect to the stated use of the model results;
   - how to use information (from evaluations covered by above guidances) as guidance for the technical user to construct a plain English characterization for the non-technical, risk manager (i.e., "model characterization language" similar to the risk characterization paradigm) to address the overall assessment with respect to the stated use of the model results (Part I).

   Research efforts could be under the auspices of the Risk Assessment Forum in consultation with the SPC/CREM and could be funded through mechanisms such as the ORD's STAR program.

3. ***Developing and utilizing a model clearinghouse to inform internal and external stakeholders on model evaluation results, availability and application experience in concert with the Program Offices and Regions.*** A possible solution for a centralized location for models is the Applications Systems Inventory (ASI) proposed by the Office of Information Resources Management (OIRM). OIRM is required to maintain an Information Systems Inventory of metadata related to EPA's computer application systems, models, modules and databases. This would require agreement by the various program offices and Regions to support such a system. At this stage, the clearinghouse is not envisioned as being resource-intensive in terms of providing technical assistance. The user of the clearinghouse would be referred to individual programs/offices for model support assistance.

4. ***Integration of developed peer review guidance and the supporting aspects of QA/QC for environmental models.*** Once the CREM has developed peer review guidance, the supporting aspects of QA/QC for environmental regulatory models (and model systems) will need to be clarified. Some of aspects of evaluation process appear

more feasible as peer review than others, i.e., evaluation of the scientific theory underlying a model versus a line by line verification of the computer code incorporating the mathematical equations or assessment of input data quality. Thus, internal Agency discussions, in consultation with the SAB, would be helpful in identifying the most appropriate areas of model development and application for peer review and those evaluations best conducted and documented as part of supporting QA/QC.

5. ***Preparing case studies (see prototypes in Appendix B) that would serve as examples of how models used in regulatory decision-making can be evaluated and the value added by such evaluations*** (e.g., Ozone Transport Assessment Group and Chesapeake Bay modeling experience like testing for the submerged aquatic vegetation ecosystem modeling framework).

6. ***Clarifying model terminology used by EPA in producing a Glossary for the "state of the art" General Guidance.*** For example, applicable current definitions for "validation," "uncertainty" for various areas of model evaluation, and "modeling error."

6.
QUESTIONS POSED BY THE COORDINATING COMMITTEE

1. How do the issues of Peer Review (external/internal) and QA/QC evaluation relate to acceptability determination?

Models are important tools supporting EPA's efforts in Risk Assessment and Management. To the extent that models erroneously estimate conditions, EPA could make costly mistakes in related decision-making. Therefore, models are covered by both QA and peer review requirements.

According to the Science Policy Council Peer Review Handbook, (EPA 100-B-98-001) models generally should be peer reviewed, and the ATFERM Guidance for Conducting External Peer Review of Environmental Regulatory Models has been incorporated in the handbook (EPA 100-B-94-001). Peer review provides an expert and independent third party review that cannot be provided by stakeholder or peer involvement and public comment. However, peer and stakeholder involvement provide valuable advice and feedback to model developers to assure a usable product (e.g., advisory panels and workshops). In 1997, EPA's Quality Assurance Division conducted an evaluation of implementation of the EPA's peer review procedures and found that few of the over 300 interviewees used the guidance on model peer review. Most, of the few who were aware of the guidance, were unclear about what was expected in implementing it and apparently had no incentive to do so because it was "only guidance."

The American National Standard "Specifications and Guidelines for Quality Systems for Environmental Data Collection and Environmental Technology Programs" (ANSI/ASQ, 1994) cited in contract and assistance agreement regulations and incorporated in the revised EPA Order 5360.1 CHG1 and QA manual 5360.1, specifies QA requirements applicable to models. The scope of the order's applicability includes "the use of environmental data collected for other purposes or from other sources (also termed secondary data), including ... from computerized data bases and information systems, results from computerized or mathematical models ..." Project level planning, implementation and assessment are addressed to assure data, whether collected or existing, "are of sufficient quantity and adequate quality for their intended use." Implementation requirements include data processing to be performed in accordance with approved instructions, methods and procedures. Also required are evaluation of new or revised software including that used for "modeling of environmental processes" and documentation of limitations on use of data (e.g., model output data).

Implementation by the SPC of the ATFERM's recommendations for the Agency to provide guidance on model selection using model acceptance criteria and information from a "Model Information System" would help to clarify what model evaluation records are needed (e.g., code verification, testing results, model selection and the application process). The model evaluation strategy proposed above could provide a process tailored to the nature of the

predictive task to be performed and the magnitude of the risk of making a wrong decision consistent with existing QA guidance. It could also clarify the complementary roles of QA and peer review tasks in model evaluation and the basis for guidance on QA Project Plans for model development and application.

The Agency's Peer Review Handbook includes under materials for peer reviewers: the charge, the work product, "associated background material" and "what is needed to complete their task." Other useful material can include "a bibliography and/or any particular relevant scientific articles from the literature." This leaves unclear what specific records are needed to adequately answer questions on model elements (EPA 100-B-94-001 Section VI) like: "What criteria were used to assess the model performance?," "What databases were used to provide an adequate test?," "What were key assumptions and the model's sensitivity to them?," "How the model performs relative to other models?," "Adequacy of documentation of model code and verification testing?," and "How well does the model report variability and uncertainty in its output?"

A number of the requirements in the Peer Review Handbook also need to be clarified:

a) What documentation is needed for peer review files of externally developed models to show "the model was independently peer reviewed with the intent of the Agency's Peer review policy for models and that EPA's proposed use" was evaluated?

b) What are the criteria needed to identify "models supporting regulatory decision making or policy/guidance of major impacts such as those having applicability to a broad spectrum of regulated entities and other stakeholders', or that will "have a narrower applicability, but with significant consequences on smaller geographic or practical scale" needing peer review?

c) What are the criteria by which decision makers judge when "a model application situation departs significantly from the situation covered in a previous peer review" so that it needs another peer review?

d) What are the criteria by which decision makers judge when "a modification of an existing adequately peer reviewed model departs significantly from its original approach" so that it needs another peer review?

e) What is the relationship of the peer review of models in the development stage often reported in journal articles (where peer review is usually performed for specific reasons for that journal and does not substitute for peer review of the Agency work product or provide accessible peer review records) and peer review addressing model use to support an Agency action?

f) What questions need to be asked in peer review of model applications supporting site specific decisions where the underlying model is "adapted to the site specific circumstances?

2. What is a consensus definition of model use acceptability criteria?

After reviewing the diversity of models and their uses across the Agency, the task group has proposed a model evaluation strategy rather than a "one size fits all" set of criteria.

3. Does acceptability correspond to a particular model, or specific applications of a model?

Specific models and applications could be accommodated in specific criteria developed by the programs.

4. Does acceptability cover only models developed by EPA or can it cover externally developed models?

Acceptability covers all models **used** in Agency regulatory decision making (see Appendix F).

5. Does acceptability mean the agency will develop a "clearinghouse" of models that meet EPA's definition of acceptable?

As discussed above it is recommended that program managers would be responsible for acceptance of models for use in their program activities. Some means of providing model evaluation status and information to potential users to be used in model selection is needed Agency-wide. The task group further recommends that a mechanism be developed for providing information responding to the model acceptance criteria to potential users to support model selection and avoid redundancy in model evaluation efforts. A number of Agency efforts might be evaluated to determine how best to leverage their resources to achieve Agency-wide goals. Some model clearinghouses already exist but often lack support. One exception is the Guideline on Air quality Models Appendices (http://www.epa.gov/ttn/scram) that provides preferred models as well as information on alternative models (e.g., regulatory use, data input, output format and options, accuracy and evaluation studies) supporting selection. As a result of the December 1997 Models 2000 Conference, an action team was formed for a Modeling Clearinghouse responding to a need perceived by the participants. OIRM is proposing to develop an Application Systems Inventory (ASI) as a repository of information about Agency software and clearinghouse to meet the Paperwork Reduction Act of 1980 and OMB Circular A-130 requirements. The ASI would integrate metadata collection requirements across the Agency and could be modified to meet specific model metadata requirements providing information on model evaluation and use. Another effort, by ORD's National Center for Environmental Assessment, is defining metadata for models that can be stored in its relational database, the Environmental Information Management System, with input through the internet and retrieval through a search engine using specific queries. Information found useful for model selection such as those listed in the Nonpoint Source Model Review example (Appendix F) is being considered for data elements. In addition, a strategy for communication needs to be developed for the public and others, like model users, to provide feedback to the EPA, possibly through the Internet at sites providing information on models and their evaluation for EPA use.

6. Would each program/region develop their own system for evaluating acceptability?

Yes, in terms of program specifications (both qualitative and quantitative).

7. Should EPA apply a generic set of criteria across the board to all categories of ERMs or should acceptability criteria differ depending on the complexity and use (e.g., screening vs. detailed assessment) of a model?

Both, generic criteria on model evaluation developed by direct interaction of the SPC and CREM with tailoring of specifications (both qualitative and quantitative) done by the programs.

7.
COST AND BENEFITS DISCUSSION

The task group requested information on costs of model evaluation activities from its members, those involved in the case histories, and the Models 2000 Steering and Implementation Team. The limited interim responses (Appendix G) were distributed for comment. Evaluation activities vary in the resources required depending on their complexity, availability of in-house expertise, and whether or not costs can be leveraged with other organizations. On the low end, the EPAMMM evaluation of a screening model evaluation without site data cost about $60,000 (for an in-house expert's 0.5 full-time equivalent). On the high end, the RADM Evaluation Study cost about $17 million for field studies, $.5 million for NOAA FTEs, and $1 million for contractors to run the model in tests in their 2.5 year effort. The Air Modeling regulatory program has used 20 to 25 staff personnel over the past 20 years with extramural support of $1.5 to 2 million per year. Model performance evaluation and peer review has cost about $150 to 200,000 per model category (2 to 10 models), although the AERMOD evaluation cost somewhat less than $50,000. In the interagency MMSOILS benchmarking evaluation, EPA's portion of the cost involved scenario development of about $50,000, execution of 4 models at about $100,000 and output comparison and reporting at about $150,000. Total coding costs estimated for the IEUBK model were about $200,000 and separate test costs were not available from the contractor. The cost would depend on the language used and when the programming documentation was done, costs being higher if documentation was done late in the process (could equal half the total coding cost). At the Models 2000 conference, Bob Carsel estimated that for a ground water model, software evaluation and documentation cost about 30% of the project cost. The costs for the AIR Dispersion Model Clearinghouse and SCRAM were about 2 FTEs/GS-13 for in-house personnel for maintaining the clearinghouse and providing regional support and regional workshops. The database, MCHISRS, cost about $50,000 for contract support over few years with little upkeep for SCRAM.

Expenditure by the Agency of the costs summarized above need to be considered in light of the benefits of better documentation and communication of the strengths and weaknesses of models. If carried out the task group's recommendations would promote systematic management of model development and use within EPA by providing a basis for consistent model evaluation and peer review. The proposed model evaluation strategy would encourage sensitivity and uncertainty analyses of environmental models and their predictions as well as clarify peer review requirements. Access to the evaluation results would improve model selection and avoid redundancy in model development and evaluation. Although these benefits would involve costs in model development for evaluation, peer review and access to evaluation results but would result in better products. Likewise, the additional cost incurred by evaluation of model application would provide feedback to developers that would improve model performance.

APPENDIX A - SCIENCE POLICY COUNCIL MODEL ACCEPTANCE CRITERIA WHITE PAPER TASK GROUP MEMBERS

Linda Kirkland, Ph.D., Chair, Quality Assurance Division, NCERQA/ORD

Brenda Johnson, Region 4

Dale Hoffmeyer, OAR

Hans Allender, Ph.D., OPPT

Larry Zaragoza, OSWER

Jerry LaVeck, OW

Thomas Barnwell, Ph.D., ORD/NERL, Athens, GA

John Fowle III, Ph.D., SAB

James Rowe, Ph.D., SPC

David Ford, Ph.D., National Center for Research in Statistics and the Environment, University of Washington

APPENDIX B - MODELS EVALUATION CASE HISTORIES

1. SHADE-HSPF Case Study (Chen et al., 1996, 1998a &b)

Regulatory Niche & Purpose:

A watershed temperature simulation model was needed for targeting critical reach locations for riparian restoration and forestry best management practices development. Evaluation of attainment of stream temperature goals (water quality criteria) was emphasized.

Model Selection:

Functional selection criteria (e.g., watershed scale and continuous-based representation, stream temperature simulation module) were used to survey and evaluate existing models resulting in the Hydrologic Simulation Program-Fortran (HSPF) model being selected as the only model meeting the requirements. Further evaluation of HSPF identified limitations in two important heat budget terms.

Data Source/Development:

A stand-alone computer program, SHADE, was developed for generating solar radiation data sets with dynamic riparian shading characteristics for use as input for water temperature simulation by HSPF after it was enhanced for computing the heat conduction between water and the stream bed for complete heat balance analysis. The case study involved generating water balance information using hydrologic simulation and then computing the effective solar radiation for stream heating to simulate stream temperature dynamics with HSPF.

Existing data sources were reviewed and appropriate meteorological, stream flow, and hourly stream temperature data for model calibration and validation were located from a fish habitat restoration project in the Upper Grande Ronde (UGR) in Oregon. Most process-oriented parameters were evaluated from known watershed attributes. Other parameters were evaluated through model calibration with recorded stream flow and temperature data based upon an understanding of the study site, HSPF application guidelines, and previous HSPF studies. Topographic, hydrographic and riparian vegetation data sets for SHADE were developed with ARC/INFO GIS for 28 fully mapped segments (the mainstem river and four major tributaries) and 9 partially mapped segments (portions of nine other tributaries).

Sensitivity Analysis:

Sensitivity analysis was noted as an important technique used extensively in designing and testing hydrologic and water quality models. To evaluate the sensitivities of simulated stream temperatures to HSPF heat balance parameters (HTRCH) and SHADE parameters, one factor (model variable or parameter) was changed at a time while holding the rest of the factors constant. Absolute sensitivity coefficients representing the change in stream temperature as a result of a unit changes in each of the two groups of model parameters were calculated by the conventional factor perturbation method. Temporal (both diurnal and seasonal) and longitudinal

variations in sensitivity were noted. Riparian shading parameters in SHADE were evaluated for stream temperature calibration to verify accuracy and reliability of SHADE computations. The solar radiation factors or SRF (deemed the most critical parameter by sensitivity analysis) as well as the diurnal, seasonal, and longitudinal variations were evaluated to verify the accuracy and reliability of SHADE computations. Significant improvement between the maximum values of SRF and the measured stream values suggested a better representation of local shading conditions by the segment based SHADE computations.

Calibration/Validation/Testing:

The model sensitivities to each parameter, as well as the diurnal, seasonal, and longitudinal variations noted, provided the basis for the stream temperature calibration. To test and demonstrate the utility and accuracy of SHADE-HSPF modeling system, hydrologic and stream temperature simulations of the watershed were conducted and visually compared to plotted measured data for two summers at 27 sites. The stream temperature calibration for 1991 and validation for 1992 generally confirmed the accuracy and robustness of SHADE-HSPF modeling system. The simulated results matched the observed points reasonably well for the majority of sites (19/27). In addition, three statistical tests were run to provide coefficients of determination and efficiency and the standard error of estimate for evaluation of critical model capability. Evaluation focused on stream temperature goals for the UGR basin (e.g., summer maximum temperature and average 7-day maximum temperature) most of the absolute errors were less than 2.5.

Simulated maximum values of stream temperature, on which the riparian restoration forecasts are based, are accurate to 2.6-3.0^{o} C. Hourly simulations have approximately the same accuracy and precision. The phase, diurnal fluctuations, and day-to-day trends in stream temperatures were generally good, indirectly confirming that riparian shading can be estimated reasonably by SHADE. Compared to the 8-10^{o}C exceedances of the temperature goals under present conditions, the model accuracy of approximately 3.0^{o}C should be adequate to assess riparian restoration scenarios.

This case history shows positive elements:

- description of regulatory use and focused on criteria for model performance
- evaluated existing models based upon specific criteria to select the model for further development and testing
- evaluation and selection of existing data for use in development, calibration, validation and testing
- sensitivity analysis and discussion of uncertainties
- good discussion of data and model limitations
- results peer reviewed by ?internal ORD review and the Journal of Environmental Engineering process as published

Concerns:

- If used by EPA, the peer review for a journal (while strengthening the scientific and technical credibility of any work product) is not a substitute for Agency work product peer review as it may not cover issues and concerns the Agency would want peer reviewed to support an Agency action.

2. TRIM.FaTE Case Study

This case study is based upon the summary (EPA, 1998 EPA-452/D-98-001) of the review of current models, the conceptual approach to the Total Risk Integrated Methodology (TRIM) framework and the first TRIM module, TRIM.FaTE, and the evaluations of TRIM.FaTE prototypes (e.g., mathematical structure and algorithms) presented to the Science Advisory Board (SAB) for an advisory review.

Regulatory Niche & Purpose:

OAR needed a multi-media, time series simulation **modeling system** to estimate multi-media impacts of both toxic and criteria air pollutants in support of Clean Air Act requirements (e.g., residual risk program, delisting petitions, urban area source program, special studies, trends, setting NAAQS, and input to regulatory impact analyses).

Model Selection:

Four multimedia, multipathway models and approaches were evaluated on the criteria that the tools have capabilities of 1) multimedia assessment; 2) ecosystem risk and exposure modeling; 3) multi-pollutant assessment; 4) addressing uncertainty and variability; and 5) accessibility and usability for EPA, states, local agencies and other stakeholders. Hazardous air pollutant (HAP) exposure and risk models also needed to adequately estimate temporal and spatial patterns of exposures while maintaining mass balance. Current criteria air pollutant models have this capability for the inhalation exposure pathway. The importance of capabilities to model pollutant uptakes, biokinetics, and dose-response for HAPs and criteria pollutants was also considered. It was found that risk and exposure assessment models, or a set of models, with the capabilities needed to address the broad range of pollutants and environmental fate and transport processes for OAQPS risk evaluations do not exist. Therefore, development of the modular TRIM framework to have varying time steps and sufficient spatial detail at varying scales, true "mass-conserving" results, transparency to support use in a regulatory context, and a truly coupled multimedia structure was begun.

Data Source/Development:

An object-oriented architecture using Visual Basic 5 application environment imbedded within Excel 97 to model the hierarchy of components of TRIM.FaTE, with a preliminary algorithm library utilizing this coding architecture, was implemented for the TRIM.FaTE prototype. The final TRIM computer framework is being designed. Where possible, existing models, tools, and databases will be adopted, necessitating their evaluation.

TRIM is planned to be a dynamic modeling system that tracks the movement of pollutant mass through a comprehensive system of compartments (e.g., physical and biological), providing an inventory of a pollutant throughout the entire system. The TRIM design is modular and, depending on the user's need for a particular assessment, one or more of six planned modules may be employed (e.g., exposure event as well as pollutant movement, uptake, biokinetics, dose response, and risk characterization). Receptors move through the compartments for

estimation of exposure. Uptake, biokinetics, and dose response models may be used to determine dose and health impacts.

Also, the TRIM.FaTE module allows flexibility to provide simulations needed for a broad range of risk evaluations because it can be formulated at different spatial and temporal scales through user selections from an algorithm library or added algorithms. The unified approach to mass transfer allows the user to change mass transfer relationships among compartments without creating a new program. This scenario differs significantly from routine application of stable single medium model programs. The mathematical linking enables a degree of precision not achieved by other models, while providing full accounting for all of the chemical mass that enters or leaves the environmental system. An outline was provided for the future user's manual for SAB's review.

Sensitivity Analysis:

Tiered sensitivity and uncertainty analyses will be integrated into the TRIM framework. All inputs to TRIM will be designed such that parameter inputs can be entered in parameter tables as default values or value distributions. Capability to estimate variability and uncertainty will be an integral part of TRIM.FaTE. Currently, only simplified sensitivity analyses have been conducted by considering the range of uncertainty in the parameter value and the linear elasticity of predicted organism concentration with respect to each input parameter. Sensitivity scores were calculated for all inputs and the sensitivity to change determined for chemical concentrations in a carnivorous fish, macrophytes, a vole, a chickadee, and a hawk with B(a)P in a steady state. Parameters with both relatively high sensitivity and a large range of uncertainty were identified and efforts focused on decreasing uncertainty that would produce the largest improvement in decreasing output uncertainty. Limitations in reliability were noted to come from relevance and availability of data to address uncertainties (e.g., about soil partition processes).

Calibration/Validation/Testing:

Four prototypes of increasing complexity were developed and evaluated. Algorithm generalizations, conceptualizations of domains (e.g., soil, groundwater, air, plant, terrestrial food chain transport), and code and data structures were provided for evaluation by the SAB panel. Also, software, routines, the databases consulted, and the data tables sources were documented and the quality of the data (e.g., distributional data for terrestrial wildlife) was discussed in comments. The TRIM.FaTE prototypes were applied to the simulation of B(a)P and phenanthrene releases in a realistic aluminum smelter test case and evaluated by comparison of the distribution of mass in multiple environmental media with results from two widely used multimedia models. TRIM.FaTE yielded results similar to the other models for some media but different results for others based upon different algorithms. Without actual measured concentrations in a controlled system, it cannot be determined which model accurately reflects reality. Limited model verification has been performed to date and more is needed.

- This is an example of peer review done early and, as planned, often.
- The elements in the ATFERM "Guidance for Conducting External Peer Review of Environmental Regulatory Models" were addressed in the information provided to the SAB panel even though this was an evaluation done when only about half of the first module was completed.
- The user's choice of algorithms needs to be an informed choice based upon evaluation of information provided in the algorithm library (e.g., requirements, assumptions). Also, documentation of the rationale for the choices, the choices (e.g., problem definition, specifications of links between data and chosen algorithms, run set up and performance), and the results need to be self-documenting to provide defensibility of the model output.
- Acquisition of data and testing of the future system needs to be carefully planned and the results similarly documented for peer review and users (e.g., limitations on use).

3. MM SOILS Case Study (Laniak, et al., 1997)

Regulatory Niche & Purpose:

EPA develops, implements, and enforces regulations that protect human and ecological health from both chemical and non-chemical stressors. EPA (like DOE) has a need to understand environmental processes that collectively release, transform and transport contaminants resulting in exposure and the probability of deleterious health effects and uses simulation models to assess exposures and risks at facilities in support of its decision making processes. MMSOILS is a multimedia model used by EPA for assessing human exposure and risk resulting from release of hazardous chemicals and radionuclides. It provides screening level analyses of potential exposure and risks and site-specific predictive assessments of exposures and risks.

Model Selection:

EPA and DOE developed a technical approach, benchmarking, to provide a comprehensive and quantitative comparison of the technical formulation and performance characteristics of three similar analytical multimedia models: EPA's MMSOILS and DOE's RESRAD and MEPAS.

Calibration/Validation/Testing:

Model design, formulation and function were examined by applying the models to a series of hypothetical problems. In comparing structure and performance of the three models, the individual model components were first isolated (e.g., fate and transport for each environmental medium, surface hydrology, and exposure/risk) and compared for similar problem scenarios including objectives, inputs, contaminants, and model endpoints. Also the integrated multimedia release, fate, transport, exposure, and risk assessment capabilities were compared.

For example, the fate and transport evaluation used a series of direct release scenarios including a specific time-series flux of contaminant from the source to the 1) atmosphere, 2) vadose zone, 3) saturated zone, and 4) surface water (river). Model estimates of contaminant concentrations at specific receptor locations were compared.

To compare the performance of all components functioning simultaneously, a hypothetical problem involving multimedia release of contaminants from a landfill was simulated. The manner and degree that individual differences in model formulation propagate through the sequence of steps in estimating exposure and risk were evaluated by comparing endpoints of concentration-based model outputs (e.g., contaminant concentrations and fluxes for each medium, time to peak concentration) as well as medium-specific and cumulative dose/risk estimates.

The results discussed differences in 1) capabilities (e.g., RESRAD and MEPAS simulate formation, decay, and transport of radionuclide decay products but MMSOILS does not); 2) constructs with respect to simulating direct releases to the various media (e.g., MMSOILS allows for varying source release constructs but does not allow for specific media to be isolated per simulation because all model modules must be executed in a simulation); 3) direct releases to the atmosphere, vadose zone, saturated zone, and surface water (e.g., all models use nearly identical formulations for transport and dispersion, resulting in close agreement with respect to airborne concentration predictions at distances greater than one kilometer from the source); 4) how surface hydrology is handled (e.g., MMSOILS does not distinguish between forms of precipitation like rainfall and snow fall); 5) direct biosphere exposure and risk (e.g., models are in complete agreement for the vast majority of methods to calculate exposure and risk with differences occurring in test scenarios for irrigation, external radiation ad dermal adsorption in contact with waterborne contamination) ; and 6) the multimedia scenario (e.g., predictions of total methylene chloride mass that volatilizes differ by a factor of 10 between the models). Results showed that the models differ with respect to 1) environmental processes included, and 2) the mathematical formulation and assumptions related to the implementation of solutions.

Peer Review of the benchmarking process and results was carried out externally by the DOE Science Advisory Board in a review of the DOE Programmatic Environmental Impact Statement and the journal, “Risk Analysis.”

This case history shows positive elements:

Results provide comparable information on model design, formulation and function that can support informed selection decisions between these models.

Concerns:

Objectives for the study did not address how well the models predicted exposures and risks relative to actual monitored releases, environmental concentrations, mass exposures or health effects. Also there are no test results of the models in applications supporting specific

regulatory and remedial action assessment needs because the study was based upon the premise that the models were applicable to the types of problems for which they were typically used. Additional information would be needed for selection decisions for model application in EPA.

Cost:

Scenario Development about	$50,000
Execution of the Models	$100,000
Output comparison and write-up (Journal articles)	$150,000
Total	$300,000

4. RADM Case Study (NAPAP Report 5, 1990)

Regulatory Niche & Purpose:

The National Acid Precipitation Assessment Program (NAPAP) needed models to assess changes in sulfate in response to proposed changes in emissions of sulfur dioxide,

Model Selection:

Regional models (linear chemistry, statistical and Lagrangian) were evaluated with new NAPAP data. An operational baseline for RADM performance was compiled based upon previous operational evaluations.

Data Source/Development:

RADM was developed to assess changes in sulfate in response to proposed changes in emissions of sulfur dioxide including addressing nonlinearity between changes in emissions and changes in deposition. Earlier simpler models were regarded as unreliable because they did not capture complex photochemistry and nonlinearities inherent in the natural system (e.g., the role of oxidants and aqueous-phase transformations). Nonlinearity (fractional changes in primary pollutant emissions are not matched by proportional changes in wet deposition of its secondary product) was of concern because at the levels of control addressed (10 million metric tons) reduction of emissions by another 20% to overcome nonproportionality and achieve a target could double control costs.

Sensitivity Analysis:

Sensitivity studies were done to determine which parameterization worked best for the RADM's meteorology and transport module, its chemistry module and the integrated model (e.g., meteorology, chemistry, emissions, and boundary conditions to the relative uncertainty in 6 key species concentrations).

Calibration/Validation/Testing:

An Eulerian Model Evaluation Program (EMEP) was carried out to established the acceptability and usefulness of RADM for the 1990 NAPAP Integrated Assessment. Key to evaluation were performance evaluations including diagnostic assessments and sensitivity analyses leading to user's confidence in the model. Guidelines and procedures were incorporated into protocols focused upon the key scientific questions supporting the application (e.g., ability to replicate spatial patterns of seasonal and annual wet deposition). Data were collected to provide robust testing of the models noting that the confidence in the model would be related to the variety of situations in the model's domain tested with observational data to show a "lack of inaccuracy" in performance. Previous model testing was limited by availability of data. Therefore, data were collected for chemical species concentrations and deposition at the surface to improve "representativeness" as well as from aircraft and special chemistry sites to support diagnostic tests for model components in a 2-year field study.

Comparisons against field data were viewed as more important in identifying weaknesses than verification (the determination of consistency, completeness, and correctness of computer code) and adequacy of the model design. However, in regional modeling the disparity in scale between site measurements and the volume-averaged prediction is a source of uncertainty. It was noted that inadequate spatial resolution in the data could produce observations that did not represent spatial patterns actually present. Such difficulties in interpretation led to linking of model uncertainty with model evaluation. Model predictions were compared to uncertainty limits of interpolated (by "kriging") observations and the observed statistically significant differences were used in evaluation (e.g., bias estimates). Kriging produced an estimate of the uncertainty (expected squared error) of the interpolated points that could be used as confidence bands for the spatial patterns. Uncertainty in the observation data came from spatial resolution in the observations, small-scale variability in the air concentration fields and wet deposition and measurement error. The model simulated the patterns (from spatially averaged fields within a grid cell) which were expected to lie within the uncertainty bounds of the corresponding pattern obtained from the observations.

Two cycles of model development, evaluation, refinement and reevaluation were carried out. The evaluation process was looked upon as iterative as the model progressed through its stages of development: 1) informal testing by the developer, 2) testing by the user with diagnostic evaluation by the developer, and 3) performance evaluation in user application. The first year's data was used in the second phase and it was planned the second's data would be used in the third. Comparative evaluation was performed where model predictions from several versions of RADM and ADOM (developed for Canada) were evaluated against field observations for a 33-day period in the fall of 1988 to see if persistent and systematic biases occurred in predictions estimating the deposition from sources at receptor areas, to estimate the change in deposition resulting from a change in emissions, and to capture nonproportionality in deposition change. Because RADM simulates very complex processes, any errors in the model's representation of physical and chemical processes could bias the model's predictions. Predictions from process modules were compared for the gas-phase chemistry, the cloud scavenging, and transport modules when shortcomings in representations of individual components in the sulfur system were noted. Capabilities of simulating annual and seasonal averages with RADM and several linear models (for wet and dry deposition and on ambient

concentrations) were evaluated. At this early stage of regional model evaluation, no viable quantitative performance standards existed (e.g., how "inaccurate" it could be). The inferred performance of the models regarding source attribution, deposition change, and air concentration change was examined based upon the evaluations and bounding analysis results, and risk of that RADM could give "misguidance" for the 1990 NAPAP Integrated Assessment was assessed. To obtain a reliable estimate of how broadly RADM's predictions could range as a function of possible errors, a "bounding" technique was developed. RADM2.1 was suggested to be used for 1990's NAPAP Integrated Assessment because it did not exhibit any biases extreme enough to preclude use if the bounding technique was used and a cautious approach was taken to use of the predictions.

EMEP data were used to address issues of acceptance and standards of performance, but Phase 2 was not covered in the report. Performance over a large number and range of tests were stated as necessary to acquire the weight-of-evidence needed to interpret the results. A multidisciplinary panel provided peer involvement.

Cost:

It was estimated that EPA provided about 18.5 Million for the 2.5 year evaluation effort (17M for field studies, .5M for NOAA FTEs, and 1M for contractors to run the model in tests).

5. EPAMMM Case Study (J. Chen and M. B. Beck, 1998. EPA/600/R-98/106.)

Regulatory Niche & Purpose:

A model was needed to screen a large number of hazardous waste facility sites with potential contamination of groundwater by leachates. The objective was to rank the sites according to their risk of exposure in the absence of in situ field observations. Those predicted to be of highest risk would have priority for remediation.

Model Selection:

The EPA Multi-Media Model (EPAMMM) was evaluated as a tool for predicting the transport and fate of contaminants released from waste a disposal facility into an environment in several media (e.g., air or subsurface environment). The model contains 7 modules: the landfill unit, the unsaturated flow field, transport of solutes in the unsaturated zone, transport of solutes in the saturated zone, transport of solutes in the surface waters, an air emissions module, and an advective transport and dispersion of the contaminant in the atmosphere (Salhotra et al. 1990 and Sharp-Hansen et al. 1990). The application evaluated was the characterization of a Subtitle D facility using 3 of the modules: flow in the unsaturated zone, transport of solutes in the unsaturated zone, and transport of the solutes in the saturated zone. Analytical and semi-analytical techniques were used to solve the basic partial differential equations of fluid flow and solute transport.

Testing:

A protocol (Beck et al. 1995 and 1997) developed for evaluation of predictive exposure models was performed in a test case when no historical data were available to be matched to the simulated responses (traditional validation). Quantitative measures of model reliability were provided and summarized in a statistic that could augment more qualitative peer review.

Three groups of tests were formulated to determine model reliability. One test assessed the uncertainties surrounding the parameterization of the model that could affect its ability to distinguish between two sites under expected siting conditions. The output uncertainty, as a function of different site characteristics, was investigated to determine if a reasonable range of model parameter uncertainty would render the power of the model to discriminate between performance of containment facilities ineffective. A generic situation was simulated under different subsurface soil, hydrological and contaminant-degradation regimes and the power of the model to distinguish between the site's containment effectiveness was tested. The probability of identical values of the residual contaminant concentration (y) at the respective receptor sites for two sites with different soil and hydrological parameterizations was evaluated to see if it was less than some quantitative threshold, such as 0.01. 0.05, or 0.10.

Another test analyzed regionalized sensitivity (Spear et al. 1994) to determine which of the model's parameters were critical to the task of predicting the contaminant's concentration exceeding the action level. The model parameters were evaluated to determine which ones were key to discriminating among the predictions of (y) in various ranges of exposures. This identified the parameters that needed the best information to determine a particular percentile of the contaminant distribution concentration at the receptor site (y). The results also provided information on the redundancy of parameters in achieving the target performance of predicting a percentile concentration. The number of key and redundant parameters can indicate model quality for the screening application.

The third test provided a more global sensitivity analysis investigating the dependence of selected statistical properties of the distributions of predicted concentrations on specific parameters. That proportion of uncertainty attached to the output (y) that derives from the uncertainty in the knowledge of a given parameter was quantified. For each individual parameter, the extent of the statistical properties of the predicted distribution (mean, variance, and 95^{th} percentile) of (y) was determined varying as a function of the point estimates assumed for the parameter. The other parameters were treated as random variables within the framework of a Monte Carlo simulation. The results of the tests show a novel form of statistic (the Quality Index of the model's design) for judging the reliability of a candidate model for performing predictive exposure assessments.

This case history shows positive elements:

- Quantitative indicators of model performance provided without in situ field data.
-

Concerns:

- Detailed knowledge of the mathematical model's function, the details of the conditions assumed for the tests and the acceptable risks in answering the questions associated with application niche are required for this type of evaluation.

Cost:

About ½ an FTE (estimated about $60,000)

APPENDIX C - MODEL VALIDATION PROTOCOL

FROM: **DRAFT** July 4, 1994

MODEL VALIDATION FOR PREDICTIVE EXPOSURE ASSESSMENTS

M B Beck *
Lee A. Mulkey **
Thomas O. Barnwell **

* Warnell School of Forest Resources
University of Georgia
Athens, Georgia 30602-2152
and
Department of Civil Engineering
Imperial College
London SW7 2BU, UK

**U.S. Environmental Protection Agency
Environmental Research Laboratory
Athens, Georgia

The beginning of the White Paper was published as:

Beck, M. B., J. R. Ravetz, L.A. Mulkey, and T.O. Barnwell. 1997. On the Problem of Model Validation for Predictive Exposure Assessments. *Stochastic Hydrology and Hydraulics* 11:229-254. Springer-Verlag.

Part 5 CONCLUSIONS

It is not reasonable to equate the validity of a model with its ability to correctly predict the future "true" behavior of the system. A judgement about the validity of a model is a judgement on whether the model can perform its designated task reliably, i.e., at minimum risk of an undesirable outcome. It follows that whomsoever requires such a judgement must be in a position to define -- in sufficient detail -- both the **task** and the **undesirable outcome**.

However desirable might be the application of "objective" tests of the correspondence between the behavior of the model and the observed behavior of the system, their results establish the reliability of the model only inasmuch as the "past observations" can be equated with the "current task specification." No-one, to the best of our knowledge, has yet developed a quantitative method of adjusting the resulting test statistics to compensate for the degree to which the "current task specification" is believed to diverge from the "past observations."

This in no way denies, however, the value of these quantitative, objective tests wherever they are applicable, i.e., in what might be called "data-rich" problem situations. Indeed, there is the prospect that in due course comparable, quantitative measures of performance validity can be developed for the substantially more difficult (and arguably more critical) "data-poor" situations, in which predictions of behavior under quite novel conditions are required by the task specification.

In this concluding section, the purpose of the protocol for model validation set out below is to provide a **consistent** basis on which to conduct the debate, where necessary, on the validity of the model in performing its designated task reliably. It seeks **not** to define what will constitute a valid model in any given situation, but to establish the framework within which the process of arriving at such a judgement can be conducted. It acknowledges that no evidence in such matters is above dispute, not even the evidence of "objective" measures of performance validity, which themselves must depend on some subjectively chosen level of an acceptable (unacceptable) difference between a pair of numbers.

5.1 The Protocol

There are three aspects to forming a judgement on the validity, or otherwise, of a model for predictive exposure assessments:

(i) the nature of the predictive **task** to be performed;

(ii) the properties of the **model**; and

(iii) the magnitude of the **risk** of making a wrong decision.

For example, if the task is identical to one already studied with the same model as proposed for the present task and the risk of making a wrong decision is low, the process of coming to a judgement on the validity of the model should be relatively straightforward and brief. Ideally, it would be facilitated by readily available, quantitative evidence of model performance validity. At the other extreme, if the task is an entirely novel one, for which a novel form of model has been proposed, and the risk of making a wrong decision is high, it would be much more difficult to come to a judgement on the validity of the model. Evidence on which to base this judgement would tend to be primarily that of an expert opinion, and therefore largely of a qualitative nature.

While the depth of the enquiry and length of the process in coming to a judgement would differ in these two examples, much the same forms of evidence would need to be gathered and presented. It is important, however, to establish responsibilities for the gathering of such evidence, for only a part of it rests with the agency charged with the development of a model. In the following it has been assumed that a second, independent agency would be responsible for specification of the task and evaluation of the risk of making a wrong decision. The focus of the protocol will accordingly be on the forms of evidence required for evaluation of the model.

5.1.1 Examination of the Model's Composition

The composition of a model embraces several attributes on which evidence will need to be presented. These are as follows:

(i) **Structure**. The structure of the model is expressed by the assembly of constituent process mechanisms (or hypotheses) incorporated in the model. A constituent mechanism might be defined as "dispersion," for example, or as "predation of one species of organism by another." The need is to know the extent to which each such constituent mechanism has been used before in any previous (other) model or previous version of the given model. There might also be a need to know the relative distribution of physical, chemical and biological mechanisms so incorporated; many scientists would attach the greatest probability of universal applicability to a physical mechanism, and the smallest such probability to a biological mechanism.

(ii) **Mathematical expression of constituent hypotheses**. This is a more refined aspect of model structure. The mechanism of "bacterial degradation of a pollutant" can be represented mathematically in a variety of ways: as a first-order chemical kinetic expression, in which the rate of degradation is proportional to the concentration of the pollutant; or as, for instance, a function of the metabolism of bacteria growing according to a Monod kinetic expression.

(iii) **Number of state variables**. In most models of predictive exposure assessments the state variables will be defined as the concentrations of contaminants or biomass of organisms at various locations across the system of interest. The greater the number of state variables included in the model the less will be the degree of aggregation and approximation in simulating both the spatial and microbial (ecological) variability in the system's behavior. In the preceding example of "bacterial degradation of a pollutant," only a single state variable would be needed to characterize the approximation of first-order chemical kinetics; two -- one each for the concentrations of both the pollutant and the (assumed) single biomass of bacteria -- would be required for the constituent hypothesis of Monod kinetics. Similarly, a lake characterized as a single, homogeneous volume of water will require just one state variable for the description of pollutant concentration within such a system. Were the lake to be characterized as two sub-volumes (a hypolimnion and an epilimnion), however, two state variables would be needed to represent the resulting spatial variability of pollutant concentration.

(iv) **Number of parameters**. The model's parameters are the coefficients that appear in the mathematical expressions representing the constituent mechanisms as a function of the values of the state variables (and/or input variables). They

are quantities such as a dispersion coefficient, a first-order decay-rate constant, or a maximum specific growth-rate constant. In an ideal world all the model's parameters could be assumed to be invariant with space and time. Yet they are in truth aggregate approximations of quantities that will vary at some finer scale of resolution than catered for by the given model. For instance, the first-order decay-rate constant of pollutant degradation subsumes the behavior of a population of bacteria; a Monod half-saturation concentration may subsume the more refined mechanism of substrate inhibition of metabolism, and so on. In problems of groundwater contamination the volumes (areas) over which the parameters of the soil properties are assumed to be uniform are intertwined with this same problem of aggregation versus refinement. There is immense difficulty, however (as already noted in discussion of the concept of **articulation**), in establishing whether a model has the correct degree of complexity for its intended task.

(v) **Values of parameters**. Again, in an ideal world the values to be assigned to the model's parameters would be invariant and universally applicable to whatever the specific sector of the environment for which a predictive exposure assessment is required. In practice there will merely be successively less good approximations to this ideal, roughly in the following descending order:

(a) The parameter is associated with an (essentially) immutable law of physics and can accordingly be assigned a single, equally immutable, value;

(b) The parameter has been determined from a laboratory experiment designed to assess a single constituent mechanism, such as pollutant biodegradation, under the assumption that no other mechanisms are acting upon the destruction, transformation, or redistribution of the pollutant within the experiment;

(c) The parameter has been determined by calibration of the model with a set of observations of the field system;

(d) A value has been assigned to the parameter on the basis of values quoted in the literature from the application of models incorporating the same mathematical expression of the same constituent process mechanism.

It is misleading to suppose that the result of (b) will be independent of an assumed model of the behavior observed in the laboratory experiment. The coefficient itself is not observed. Instead, for example, the concentration of pollutant remaining undegraded in the laboratory beaker or chemostat is

observed. Once a mathematical description of the mechanism assumed to be operative in the experiment is postulated, then the value of the parameter can be inferred from matching the performance of this model with the observations (which in effect is the same procedure as that of (c).

(vi) **Parameter uncertainty**. Evidence should be presented on the range of values assigned to a particular parameter in past studies and/or on the magnitude and (where available) statistical properties of the estimation errors associated with these values. In many cases it might be sufficient to assume that such ranges of values and distributions of errors are statistically independent of each other, but this can be misleading. Supplementary evidence of the absence/presence of correlation among the parameter estimates and errors could be both desirable and material to the judgement on model validity. For example, unless determined strictly independently -- and it is not easy to see how that might be achieved -- the values quoted for a bacterial growth-rate constant and death-rate constant are likely to be correlated. A pair of low values for both parameters can give the same net rate of growth as a pair of high values, and knowledge of such correlation can influence both the computation of, and assessment of, the uncertainty attaching to a prediction of future behavior.

(vii) **Analysis of parameter sensitivity.** The extent to which the predictions of the model will change as a result of alternative assumptions about the values of the constituent parameters can be established from an analysis of parameter sensitivity. On its own such information provides only a weak index of model validity. It may be used, nevertheless, to supplement a judgement on the model's compositional validity based on the foregoing categories of evidence. In the absence of any knowledge of parameter uncertainty an analysis of sensitivity may yield insight into the validity of the model's composition through the identification, in extreme cases, of those "infeasible" values of the parameters that lead to unstable or absurd predictions. It could be used thus to establish in crude terms the domain of applicability of the model, i.e., ranges of values for the model's parameters for which "sensible" behavior of the model is guaranteed. In the presence of information on parameter uncertainty an analysis of sensitivity may enable rather more refined conclusions about the validity of the model. In particular, a highly sensitive, but highly uncertain, parameter is suggestive of an ill-composed model.

It is clearly impossible to divorce an assessment of the evidence on the model's compositional validity -- its intrinsic properties and attributes -- from the current task specification. In particular, the less immutable the hypothesis (law) incorporating a given parameter is believed to be, the more relevant will become a judgement about the degree to which the current task specification deviates from those under which the values previously quoted for this parameter were derived. Such judgement will be especially difficult to make in the case of quantifying the correspondence (or divergence) between the laboratory conditions used to determine a rate constant and

the field conditions for which a predictive exposure assessment is required. The judgement, nevertheless, is directed at the internal composition of the model, albeit conditioned upon the degree of similarity between the current and previous task definitions.

5.1.2 Examination of the Model's Performance

Evidence must also be assembled from the results of tests of a model's performance against an external reference definition of the prototype (field) system's behavior. This will have various levels of refinement, approximately in the following ascending order.

(i) **Unpaired tests**. In these the coincidence between values for the model's state variables and values observed for corresponding variables of the prototype system at identical points in time and space is of no consequence. It is sufficient merely for certain aggregate measures of the collection of model predictions and the collection of field data to be judged to be coincident. For example, it might be required that the mean of the computed concentrations of a contaminant in a representative (model) pond over an annual cycle is the same as the mean of a set of observed values sampled on a casual, irregular basis from several ponds in a geologically homogeneous region. Within such unpaired tests, there are further, subsidiary levels of refinement. A match of mean values alone is less reassuring than a match of both the means and variances, which is itself a less incisive test than establishing the similarity between the two entire distributions.

(ii) **Paired tests**. For these it is of central concern that the predictions from the model match the observed values at the same points in time and space. Again, as with the unpaired tests, subsidiary levels of refinement are possible, in providing an increasingly comprehensive collection of statistical properties for the errors of mismatch so determined.

(iii) **Sequence of errors**. A paired-sample test, as defined above, makes no reference to the pattern of the errors of mismatch as they occur in sequence from one point in time (or space) to the next. When sufficient observations are available a test of the temporal (or spatial) correlations in the error sequences may yield strong evidence with which to establish the performance validity of the model. In this case a "sufficiency" of data implies observations of the contaminant concentration at frequent, regular intervals over relatively long, unbroken periods.

In much the same way as it is not possible to divorce an assessment of the compositional validity of a model from its current and past task specifications, so it is not possible to divorce an assessment of performance validity from the composition of the model. Thus a further two categories of evidence are relevant.

(iv) **Calibration**. The task of model calibration necessarily involves adjustment and adaptation of the model's composition. The extent to which the values of the model's parameters have thereby been altered in order for the model to fit the calibration data set may render inadmissible the use of any associated error statistics for the purposes of judging model validity. It is therefore especially relevant for evidence of this form to be declared.

(v) **Prediction uncertainty**. All models may be subjected to an analysis of the uncertainty attaching to their predictions. Such an analysis will depend on the composition of the model -- through the quantification of parameter uncertainty; and it will depend upon the task specification, through a statement of the scenarios for the input disturbances and initial state of the system, i.e., the boundary and initial conditions for the solution of the model equations. The fact that the ambient concentration of the contaminant cannot be predicted with sufficient confidence does not necessarily signify an invalid model, however. For there are three sources of uncertainty in the predictions, two of which (the initial and boundary conditions) are independent of the model. Good practice in the analysis of prediction uncertainty (if a judgement on model validity is the objective) should therefore include some form of ranking of the contributions each source of uncertainty makes to the overall uncertainty of the prediction. Where Monte Carlo simulation is used to compute the distributions of the uncertain predictions, some -- perhaps many -- runs of the model may fail to be completed because of combinations of the model's parameter values leading to unstable or absurd output responses. As with an analysis of sensitivity, this provides useful information about the robustness of the model and restrictions on its domain of applicability. The less the model is found to be restricted, so the greater is the belief in its validity. In some cases, it may be feasible and desirable to state the output responses expected of the model in order for the task specification to be met, thus enabling a more refined assessment of the domain of applicability of the model (as in discussion of the concept of **relevance**). The use of combinations of parameter values leading to unacceptable deviations from the behavior of the task specification can be placed under restrictions.

5.1.3 Task specification

Judgements on both the compositional and performance validity of the model are inextricably linked with an assessment of the extent to which the current task

specification diverges from the task specifications of previous applications of the model. Categories of evidence relating to the fundamental properties of the task specification must therefore be defined, in a manner similar to those assembled in order to conduct an assessment of the model.

For example, a model used previously for prediction of a chronic exposure at a single site with homogeneous environmental properties may well not be valid -- in terms of performing its task reliably -- for the prediction of an acute exposure at several sites with highly heterogeneous properties. It is not that the model is inherently incapable of making such predictions, but that there is an element of extrapolation into novel conditions implied by the different task specification. It is not the purpose of this document, however, to provide anything other than a very preliminary indication of the categories of evidence required to assess the degree of difference between current and past task specifications, as follows.

(i) **The contaminants**. The class(es) of chemicals into which the contaminant would most probably fall, such as chlorinated hydrocarbon, or aromatic compound, for example, must be specified. The number of such chemicals to be released, and their interactions (synergism, antagonism, and so on) vis a vis the state variables of interest in the environment, must also be specified.

(ii) **The environment**. Several attributes can be employed to characterize the similarities and differences among the environments into which the contaminant is to be released. These include, inter alia, the geological, hydrological, and ecological properties of the sites of interest, together with statements of the homogeneity, or heterogeneity, of the site according to these attributes.

(iii) **Target organism, or organ.**

(iv) **Nature of exposure.** The obvious distinction to be made in this case is between acute and chronic exposures of the target organism to the contaminant.

APPENDIX D - REQUIREMENTS FOR A "STATE OF THE ART" EVALUATION PROCESS

Changes in the Concept of Model Validation

Over the past 10 years previous model validation approaches have been recognized as unsatisfactory. It is no longer accepted that models can be validated as defined by ASTM standard E 978-84 (e.g., comparison of model results with numerical data independently derived from experience or observation of the environment) and then considered to be "true". The discussions supporting the idea that models could not be validated (Konikow and Bredehoeft 1992, Oreskes et al. 1994), focused on hydrological models. Typically these models are first calibrated, i.e., parameter values are estimated using one data set, and then the effectiveness of that parameterization is examined using a second, independent data set. However, one, or even more, successful tests using particular data sets does not mean that a model is valid in the sense of being *true* and able to make reliable predictions for unknown future conditions. The realization of this problem has led ASTM to update its definition of model validation to: "**a test** of the model with known input and output information that is used **to assess that the calibration parameters are accurate** without further change" (ASTM E 978 - 92).

Practical approaches to validation have varied between environmental and ecological modelers. In ecology both the objectives of modeling and the data available for calibration and testing have frequently differed from those used in environmental modeling, such as hydrology. The objective of ecosystem modeling, as a particular example, has been to synthesize information about an ecosystem from a range of sources and no integrated calibration for the whole model may be possible. In this way an ecosystem model represents a complex ecological theory and there may be no independent data set to provide a test of the complete model. Consequently, the inability to validate models, in the sense of them being considered as absolutely true, has been at least tacitly accepted in ecosystem modeling for some time. Recent developments in environmental models, such as TRIM.FaTE and other multi-media type models, are similar in their methods of construction to ecosystem models. An approach for such models is **to replace validation, as though it were an endpoint** that a model could achieve, **with model evaluation as a process** that examines each of the different elements of theory, mathematical construction, software construction, calibration and testing with data.

APPENDIX E - TYPES OF MODELS USED BY EPA

Physical Models:

Atmospheric Emissions Models (e.g., GloED)
Water Treatment/Distribution Models (e.g., EPANET)
Emissions Control Models (e.g., IAPCS)
Stream Flow Model (e.g., PC-DFLOW)
Atmospheric Models (e.g., Models-3, ISC)
Chemical Estimation Models (e.g., AMEM)
Subsurface (e.g., SESOIL/AT123D)
Surface Water (e.g., SED3D,)

Biological Models:

Atmospheric Models (e.g., BEIS-2)
Chemical Estimation Models (e.g., BIOWIN)
Ecological Exposure Models (e.g., ReachScan, FGETS)
Human Health Models (e.g., TherdbASE, DEEM)
Subsurface (e.g., PRZM2)

Multimedia Models:

(MULTIMED, MULTIMDP, GEMS, PCGEMS, MMSOIL, SEAS, TRIM, HWIR, MIMS)

APPENDIX F - NONPOINT SOURCE MODEL REVIEW EXAMPLE

This is an example report showing the type of model information used in the 1991 review (Donigian and Huber, 1991).

1. Name of the Method

Hydrological Simulation Program—Fortran (HSPF)

Stream Transport and Agricultural Runoff of Pesticides for Exposure Assessment (STREAM)

2. Type of Method

___	Surface Water Model:	Simple Approach
xxx	Surface Water Model:	Refined Approach
___	Air Model:	Simple Approach
___	Air Model:	Refined Approach
___	Soil (Groundwater) Model:	Simple Approach
xxx	Soil (Groundwater) Model:	Refined Approach
___	Multi-media Model:	Simple Approach
___	Multi-media Model:	Refined Approach

3. Purpose/Scope

Purpose: Predict concentrations of contaminants in

xxx	Runoff Waters
xxx	Surface Waters
xxx	Ground Waters

Source/Release Types:

xxx	Continuous	xxx	Intermittent		
xxx	Single	xxx	Multiple	xxx	Diffuse

Level of Application:

xxx Screening xxx Intermediate xxx Detailed

Type of Chemicals:

xxx	Conventional	xxx	Organic	___	Metals

Unique Features:

- xxx Addresses Degradation Products
- xxx Integral Database/Database Manager
- ___ Integral Uncertainty Analysis Capabilities
- xxx Interactive Input/Execution Manager

4. Level of Effort

System setup:	xx	mandays	xx	manweeks	__	manmonths	__	manyear
Assessments:	__	mandays	xx	manweeks	xx	manmonths	__	manyear

(Estimates reflect order-of-magnitude values and depend heavily on the experience and ability of the assessor.)

5. Description of the Method/Techniques

Hydrological Simulation Program—FORTRAN (HSPF) is a comprehensive package for simulation of watershed hydrology and water quality for both conventional and toxic organic pollutants. HSPF incorporates the watershed scale ARM and NPS models into a basic-scale analysis framework that includes fate and transport in one-dimensional stream channels. It is the only comprehensive model of watershed hydrology and water quality that allows the integrated simulation of land and soil contaminant runoff processes with instream hydraulic and sediment-chemical interactions.

The result of this simulation is a time history of the runoff flow rate, sediment load, and nutrient and pesticide concentrations, along with a time history of water quantity and quality at any point in a watershed. HSPF simulates three sediment types (sand, silt, and clay) in addition to a single organic chemical and transformation products of that chemical. The transfer and reaction processes included are hydrolysis, oxidation, photolysis, biodegradation, a volatilization, and sorption. Sorption is modeled as a first-order kinetic process in which the user must specify a desorption rate and an equilibrium partition coefficient for each of the three solid types. Resuspension and settling of silts and clays (cohesive solids) are defined in terms of shear stress at the sediment-water interface. For sands, the capacity of the system to transport sand at a particular flow is calculated and resuspension or settling is defined by the difference between

the sand in suspension and the capacity. Calibration of the model requires data for each of the three solids types. Benthic exchange is modeled as sorption/desorption and desorption/scour with surficial benthic sediments. Underlying sediment and pore water are not modeled.

6. Data Needs/Availability

Data needs for HSPF are extensive. HSPF is a continuous simulation program and requires continuous data to drive the simulations. As a minimum, continuous rainfall records are required to drive the runoff model and additional records of evapotranspiration, temperature, and solar intensity are

desirable. A large number of model parameters can also be specified although default values are provided where reasonable values are available. HSPF is a general-purpose program and special attention has been paid to cases where input parameters are omitted.

Option flags allow bypassing of whole sections of the program where data are not available.

7. Output of the Assessment

HSPF produces a time history of the runoff flow rate, sediment load, and nutrient and pesticide concentrations, along with a time history of water quantity and quality at any point in a watershed. Simulation results can be processed through a frequency and duration analysis routine that produces output compatible with conventional toxicological measures (e.g., 96-hour LC50).

8. Limitations

HSPF assumes that the "Stanford Watershed Model" hydrologic model is appropriate for the area being modeled. Further, the instream model assumes the receiving water body model is well-mixed with width and depth and is thus limited to well-mixed rivers and reservoirs. Application of this methodology generally requires a team effort because of its comprehensive nature.

9. Hardware/Software Requirements

The program is written in standard FORTRAN 77 and has been installed on systems as small as IBM PC/AT-compatibles. A hard disk is required for operation of the program and a math co-processor is highly recommended. No special peripherals other than a printer are required. The program is maintained for both the IBM PC-compatible and the DEC/VAX with VMS operating system. Executable code prepared with the Ryan-McFarland FORTRAN compiler and PLINK86 linkage editor is available for the MS/DOS environment. Source code only is available for the VAX environment.

The program can be obtained in either floppy disk format for MS/DOS operation systems or on a 9-TRK magnetic tape with installation instructions for the DEC VAX VMS environment. This program has been installed on a wide range of computers world-wide with no or minor modifications.

10. Experience

HSPF and the earlier models from which it was developed have been extensively applied in a wide variety of hydrologic and water quality studies (Barnwell and Johanson, 1981; Barnwell and Kittle, 1984) including pesticide runoff testing (Lorber and Mulkey, 1981), aquatic fate and transport model testing (Mulkey et al., 1986; Schnoor et al., 1987) analyses of agricultural best management practices (Donigian et al., 1983a; 1983b; Imhoff et al., 1983) and as part of pesticide exposure assessments in surface waters (Mulkey and Donigian, 1984).

An application of HSPF to five agricultural watersheds in a screening methodology for pesticide review is given in Donigian (1986). The Stream Transport and Agricultural Runoff for Exposure Assessment (STREAM) Methodology applies the HSPF program to various test watersheds for five major crops in four agricultural regions in the U.S., defines a "representative watershed" based on regional conditions and an extrapolation of the calibration for the test watershed, and performs a sensitivity analysis on key pesticide parameters to generate cumulative frequency distributions of pesticide loads and concentrations in each region. The resulting methodology requires the user to evaluate only the crops and regions of interest, the pesticide application rate, and three pesticide parameters—the partition coefficient, the soil/sediment decay rate, and the solution decay rate.

11. Validation/Review

The program has been validated with both field data and model experiments and has been reviewed by independent experts. Numerous citations for model applications are included in the References below. Recently, model refinements for instream algorithms related to pH and sediment-nutrient interactions have been sponsored by the USGS and the EPA Chesapeake Bay Program, respectively.

12. Contact

The model is available from the Center for Exposure Assessment Modeling at no charge. Mainframe versions of the programs compatible with the DEC VAX systems are available on standard one-half inch, 9-track magnetic tape. When ordering tapes, please specify the type of computer system that the model will be installed on (VAX, PRIME, HP, Cyber, IBM, etc.), whether the tape should be non-labeled (if non-labeled specify the storage format, EBCDIC or ASCII), or if the tape should be formatted as a VAX files-11, labeled (ASCII) tape for DEC systems. Model distributions tapes contain documentation covering installation instructions on DEC systems, FORTRAN source code files, and test input data sets and output files that may be used to test and confirm the installation of the model on your system. Users are responsible for installing programs.

Requests for PC versions of the models should be accompanied by 8 formatted double-sided, double-density (DS/DD), error-free diskettes. Please do not send high-density (DD/HD) diskettes. Model distribution diskettes contain documentation covering installation instructions on PC systems, DOS batch files for compiling, linking, and executing the model, executable

task image(s) ready for execution of the model(s), all associated runtime files, and test input data sets and corresponding output files that may be used to test and confirm the installation of the model on your PC or compatible system.

To obtain copies of the models, please send 9-track specifications or the appropriate number of formatted diskettes to the attention of David Disney at the following address:

Center for Exposure Assessment Modeling
U.S. Environmental Protection Agency
Environmental Research Laboratory
Athens, Georgia 30613
(404) 546-3123
USA

Program and/or user documentation, or instructions on how to order documentation, will accompany each response.

13. References

Barnwell, T.O. 1980. An Overview of the Hydrologic Simulation Program—FORTRAN, a Simulation Model for Chemical Transport and Aquatic Risk Assessment. *Aquatic Toxicology and Hazard Assessment: Proceedings of the Fifth Annual Symposium on Aquatic Toxicology,* ASTM Special Tech. Pub. 766, ASTM, 1916 Race Street, Philadelphia, PA 19103.

Barnwell, T.O. and R. Johanson. 1981. HSPF: A Comprehensive Package for Simulation of Watershed Hydrology and Water Quality. *Nonpoint Pollution Control: Tools and Techniques for the Future*. Interstate Commission on the Potomac River Basin, 1055 First Street, Rockville, MD 20850.

Barnwell, T.O. and J.L. Kittle. 1984. Hydrologic Simulation Program—FORTRAN: Development, Maintenance and Applications. *Proceedings Third International Conference on Urban Storm Drainage*. Chalmers Institute of Technology, Goteborg, Sweden.

Bicknell, B.R., A.S. Donigian Jr. and T.O. Barnwell. 1984. Modeling Water Quality and the Effects of Best Management Practices in the Iowa River Basin. *J. Wat. Sci. Tech.* 17:1141-1153.

Chew, Y.C., L.W. Moore, and R.H. Smith. 1991. Hydrologic SIMULATION of Tennessee's North Reelfoot Creek Watershed. *J. Water Pollution Control Federation* 63(1):10-16.

Donigian, A.S., Jr., J.C. Imhoff and B.R. Bicknell. 1983. *Modeling Water Quality and the Effects of Best Management Practices in Four Mile Creek, Iowa*. EPA Contract No. 68-03-2895, Environmental Research Laboratory, U.S. EPA, Athens, GA 30613.

Donigian, A.S., Jr., J.C. Imhoff, B.R. Bicknell and J.L. Kittle, Jr. 1984. *Application Guide for the Hydrological Simulation Program—FORTRAN EPA*. 600/3-84-066, Environmental Research Laboratory, U.S. EPA, Athens, GA. 30613.

Donigian, A.S., Jr., D.W. Meier and P.P. Jowise. 1986. *Stream Transport and Agricultural Runoff for Exposure Assessment: A Methodology*. EPA/600/3-86-011, Environmental Research Laboratory, U.S. EPA, Athens, GA 30613.

Donigian, A.S., Jr., B.R. Bicknell, L.C. Linker, J. Hannawald, C. Chang, and R. Reynolds. 1990. *Chesapeake Bay Program Watershed Model Application to Calculate Bay Nutrient Loadings: Preliminary Phase I Findings and Recommendations*. Prepared by AQUA TERRA Consultants for U.S. EPA Chesapeake Bay Program, Annapolis, MD.

Hicks, C.N., W.C. Huber and J.P. Heaney. 1985. Simulation of Possible Effects of Deep Pumping on Surface Hydrology Using HSPF. *Proceedings of Stormwater and Water Quality Model User Group Meeting*. January 31–February 1, 1985. T.O. Barnwell, Jr., ed. EPA-600/9-85/016. Environmental Research Laboratory, Athens, GA.

Johanson, R.C., J.C. Imhoff, J.L. Kittle, Jr. and A.S. Donigian. 1984. *Hydrological Simulation Program—FORTRAN (HSPF): Users Manual for Release 8.0*. EPA-600/3-84-066, Environmental Research Laboratory, U.S. EPA, Athens, GA. 30613.

Johanson, R.C. 1989. Application of the HSPF Model to Water Management in Africa. *Proceedings of Stormwater and Water Quality Model Users Group Meeting.* October 3-4, 1988. Guo, et al., eds. EPA-600/9-89/001. Environmental Research Laboratory, Athens, GA.

Lorber, M.N. and L.A. Mulkey. 1982. An Evaluation of Three Pesticide Runoff Loading Models. *J. Environ. Qual.* 11:519-529.

Moore, L.W., H. Matheny, T. Tyree, D. Sabatini and S.J. Klaine. 1988. Agricultural Runoff Modeling in a Small West Tennessee Watershed. *J. Water Pollution Control Federation* 60(2):242-249.

Motta, D.J. and M.S. Cheng. 1987. The Henson Creek Watershed Study. *Proceedings of Stormwater and Water Quality Users Group Meeting.* October 15-16, 1987. H.C. Torno, ed. Charles Howard and Assoc., Victoria, BC, Canada.

Mulkey, L.A., R.B. Ambrose, and T.O. Barnwell. 1986. Aquatic Fate and Transport Modeling Techniques for Predicting Environmental Exposure to Organic Pesticides and Other Toxicants—A Comparative Study. *Urban Runoff Pollution,* Springer-Verlag, New York, NY.

Nichols, J.C. and M.P. Timpe. 1985. Use of HSPF to Simulate Dynamics of Phosphorus in Floodplain Wetlands over a Wide Range of Hydrologic Regimes. *Proceedings of Stormwater and Water Quality Model Users Group Meeting*. January 31–February 1, 1985. T.O. Barnwell, Jr., ed. EPA-600/9-85/016, Environmental Research Laboratory, Athens, GA.

Schnoor, J.L., C. Sato, D. McKetchnie, and D. Sahoo. 1987. *Processes, Coefficients, and Models for Simulating Toxic Organics and Heavy Metals in Surface Waters*. EPA/600/3-87/015. U.S. Environmental Protection Agency, Athens, GA 30613.

Schueler, T.R. 1983. *Seneca Creek Watershed Management Study, Final Report, Volumes I and II*. Metropolitan Washington Council of Governments, Washington, DC.

Song, J.A., G.F. Rawl, and W.R. Howard. 1983. Lake Manatee Watershed Water Resources Evaluation using Hydrologic Simulation Program—FORTRAN (HSPF). *Colloque sur la Modelisation des Eaux Pluviales*. Septembre 8-9, 1983. P. Beron, et al., T. Barnwell, editeurs. GREMU—83/03 Ecole Polytechnique de Montreal, Quebec, Canada.

Sullivan, M.P. and T.R. Schueler. 1982. The Piscataway Creek Watershed Model: A Stormwater and Nonpoint Source Management Tool. *Proceedings Stormwater and Water Quality Management Modeling and SWMM Users Group Meeting*. October 18-19, 1982. Paul E. Wisner, ed. Univ. of Ottawa, Dept. Civil Engr., Ottawa, Ont., Canada.

Weatherbe, D.G. and Z. Novak. 1985. Development of Water Management Strategy for the Humber River. *Proceedings Conference on Stormwater and Water Quality Management Modeling*. September 6-7, 1984. E.M. and W. James, ed. Computational Hydraulics Group, McMaster University, Hamilton, Ont., Canada.

Woodruff, D.A., D.R. Gaboury, R.J. Hughto and G.K. Young. 1981. Calibration of Pesticide Behavior on a Georgia Agricultural Watershed Using HSP-F. *Proceedings Stormwater and Water Quality Model Users Group Meeting*. September 28-29, 1981. W. James, ed. Computational Hydraulics Group, McMaster University, Hamilton, Ont., Canada.

Udhiri, S., M-S Cheng and R.L. Powell. 1985. The Impact of Snow Addition on Watershed Analysis Using HSPF. *Proceedings of Stormwater and Water Quality Model Users Group Meeting*. January 31–February 1, 1985. T.O. Barnwell, Jr., ed. EPA-600/9-85/016, Environmental Research Laboratory, Athens, GA.

APPENDIX G - COST ESTIMATES

The following are the cost estimates received in response to our inquiries:

MMSOILS Benchmarking Evaluation (EPA's portion, DOE unknown)

Cost: Scenario Development about $50,000; Execution of 4 models about $100,000; Output comparison and write-up (Journal articles) $150,000; Total = $300,000.

RADM Evaluation Study

Cost: It was estimated that EPA provided about 18.5 Million for the 2.5 year evaluation effort (17 M for field studies, .5M for NOAA FTEs, and 1M for contractors to run the model in tests). [files for contract support have been disposed of]

AIR Dispersion Model Clearinghouse and SCRAM:

2 FTEs/GS-13 - clearinghouse, regional support, and support of regional workshops

MCHISRS $50,000 contractor over few years with little upkeep for SCRAM

Air Modeling Regulatory program - 20 to 25 staff for 20 years with extramural support $1.5 to 2 M per year

Model performance evaluation and peer review about $150 to 200 K per model category (2-10 models)

AERMOD over 6 years exceeded $500K the evaluation portion is less than 10% of the total.

EPAMMM Evaluation (Screening model evaluation without site data)

½ FTE (about $60,000)

Software Evaluation and Documentation Costs:

Checked with Margarette Shovlin who said costs are not broken out to the level of model code testing and verification or documentation on a level of effort contract. Larry Zaragoza attempted to get estimates for the IEUBK model and found the same thing getting the information would be tedious- a special request to the ESDS's SAIC management would have to be made by OARM.

Larry estimated the IEUBK coding cost about $200K but its hard to separate out test costs and it depends on the language used and how close to the actual programming documentation is done. He estimated that if documentation was done late in the process the cost could equal half

the total project cost. At the Models 2000 conference Bob Carsel estimated that for a ground water model, software evaluation and documentation cost about 30 % of the project cost.

APPENDIX H - REFERENCES

American Society for Testing and Materials. 1992. Standard Practice for Evaluating Environmental Fate Models of Chemicals. *Standard 978-92*. Philadelphia: American Society for Testing and Materials.

Beck, M. B., L.A. Mulkey, and T.O. Barnwell. 1994. *Draft Model Validation for Predictive Exposure Assessments.* Presented to the Risk Assessment Forum, July 1994.

Beck, M. B., J. R. Ravetz, L.A. Mulkey, and T.O. Barnwell. 1997. On the Problem of Model Validation for Predictive Exposure Assessments. *Stochastic Hydrology and Hydraulics* 11: 229-254. Springer-Verlag.

Beck, M. B., L.A. Mulkey, T.O. Barnwell, and J. R. Ravetz. 1997. *Model Validation for Predicative Exposure Assessments.* 1995 International Environmental Conference Proceedings, p. 973-980. TAPPI Proceedings.

Chen, J. and M. B. Beck. 1998. *Quality Assurance of Multi-Media Model for Predictive Screening Tasks.* (EPA/600/R-98/106). August 1998.

Chen, Y. D., S.C. McCutcheon, R.F. Carsel, D.J. Norton, and J.P. Craig. 1996. Enhancement and Application of HSPF for Stream Temperature Simulation in Upper Grande Ronde Watershed, Oregon. *Watershed '96* p. 312-315.

Chen, Y. D., R.F. Carsel, S.C. McCutcheon, and W.L. Nutter. 1998. Stream Temperature Simulation of Forested Riparian Areas: I. Watershed-scale Model Development. *Journal of Environmental Engineering* p. 304-315. April 1998.

Chen, Y. D., S.C. McCutcheon, D.J. Norton, and W.L. Nutter. 1998. Stream Temperature Simulation of Forested Riparian Areas: II. Model Application. *Journal of Environmental Engineering* p. 316-328. April 1998.

Cicchetti, D.V. 1991. The Reliability of Peer Review for Manuscript and Grant Submissions: A Cross-Disciplinary Investigation. *Behavioral and Brain Sciences* 14:119-186.

Donigian, A.S., Jr. and W.C. Huber. 1991. *Modeling of Nonpoint Source Water Quality on Urban and Non-urban Areas*. EPA/600/3-91/039 (NTIS PB92-109115) U.S. Environmental Protection Agency, Athens, GA.

Doucet, P. and P.B. Sloep. 199?. *Mathematical Modeling in the Life Sciences*. Ellis Horwood, New York. p. 280-281.

Gillies, D. 1993. *Philosophy of Science in the Twentieth Century*. Blackwell, Oxford, U.K.

Konikow, L.F. and J.D. Bredehoeft. 1992. Ground-Water Models Cannot Be Validated. *Adv. Water Resources* 15:75-83.

Oreskes, N., K. Shrader-Frechette, and K. Belitz. 1994. Verification, Validation, and Confirmation of Numerical Models in the Earth Sciences. *Science* 263:641-646.

Laniak, G.F., J. G. Droppo, E. R. Faillace, E. K. Gnanapragasam, W.B. Mills, D.L. Strenge, G. Whelan, and C. Yu. 1997. An Overview of a Multimedia Benchmarking Analysis for Three Risk Assessment Models: RESRAD, MMSOILS, and MEPAS. *Risk Analysis* 17(2):1-23. April 1997.

Lee, S.B., V. Ravi, J. R. Williams, and D.S. Burden. 1996. Subsurface Fluid Flow (Groundwater and Vadose Zone) Modeling: Application of Subsurface Modeling Application. *ASTM Special Technical Publication 1288* p. 3-13.

National Acid Precipitation Assessment Program. 1990. Acidic Deposition: State of Science and Technology: Report 5. *Evaluation of Regional Acidic Deposition Models (Part I) and Selected Applications of RADM (Part II).* September 1990.

National Institute of Standards and Technology. 1996. *NIST Special Publication 500-234: Reference Information for the Software Verification and Validation Process.* (http:\\hissa.ncsl.nist.gov/HHRFdata/Artifacts/ITLdoc/234/val-proc.html)

Peters, D.P. and S.J. Ceci. 1982. Peer-Review Practices of Psychology Journals: The Fate of Published Articles Submitted Again. *Behavioral and Brain Sciences* 5:187-255.

USEPA. 1993. *Computer Models Used to Support Cleanup Decision-Making at Hazardous and Radioactive Waste Sites*. EPA 402-R-93-005, March 1993. (NTIS, PB93-183333/XAB).

USEPA. 1993. *Environmental Pathway Models–Ground-Water Modeling in Support of Remedial Decision-Making at Sites Contaminated with Radioactive Material*. EPA 402-R-93-009, March 1993. (NTIS, PB93-196657/XAB).

USEPA. 1994. *Technical Guide to Ground-Water Model Selection at Sites Contaminated with Radioactive Substances*. EPA 402-R-94-012, September 1994. (NTIS, PB94-205804/XAB).

USEPA. 1996. *Documenting Ground-Water Modeling at Sites Contaminated with Radioactive Substances*. EPA 540-R-96-003, January 1996. (NTIS, PB96-963302/XAB).

USEPA Office of Air Quality Planning & Standards. 1998. *The Total Risk Integrated Methodology - Technical Support Document for the TRIM.FaTE Module Draft.* EPA-452/D-98-001. March 1998.

USEPA Office of the Administrator. 1994. *Guidance for Conducting External Peer Review of Environmental Peer Review of Environmental Regulatory Models.* EPA 100-B-94-001. July 1994.

USEPA Office of Research and Development. 1997. *Guiding Principles for Monte Carlo Analysis.* EPA/630/R-97/001. March 1997.

USEPA Office of Science Policy, Office of Research and Development. 1998. *Science Policy Council Handbook: Peer Review.* EPA 100-B-98-001. January 1998.

USEPA Solid Waste and Emergency Response. 1994. *Report of the Agency Task Force on Environmental Regulatory Modeling: Guidance, Support Needs, Draft Criteria and Charter.* EPA 500-R-94-001. March 1994.

US General Accounting Office. 1997. *Report to the Chairman, Subcommittee on Oversight and Investigations, Committee on Commerce, House of Representatives: Air Pollution - Limitations of EPA's Motor Vehicle Emissions Model and Plans to Address Them.* GAO/RCED-97-21 September 1997.

van Vallen, L. and F.A. Pitelka. 1974. Commentary - Intellectual Censorship in Ecology. *Ecology* 55:925-926.